NORTH ATLANTIC TREATY ORGANIZATION

SCIENCE AND TECHNOLOGY ORGANIZATION

AC/323(SAS-114)TP/928

www.sto.nato.int

STO TECHNICAL REPORT **TR-SAS-114**

Assessment and Communication of Uncertainty in Intelligence to Support Decision-Making

(Évaluation et communication de l'incertitude dans le renseignement en vue de faciliter la prise de décision)

Final Report of Research Task Group SAS-114.

Edited by David R. Mandel.

Published June 2020

Distribution and Availability on Back Cover



NORTH ATLANTIC TREATY ORGANIZATION

SCIENCE AND TECHNOLOGY ORGANIZATION

AC/323(SAS-114)TP/928

www.sto.nato.int

STO TECHNICAL REPORT **TR-SAS-114**

Assessment and Communication of Uncertainty in Intelligence to Support Decision-Making

(Évaluation et communication de l'incertitude dans le renseignement en vue de faciliter la prise de décision)

Final Report of Research Task Group SAS-114.

Edited by David R. Mandel

The NATO Science and Technology Organization

Science & Technology (S&T) in the NATO context is defined as the selective and rigorous generation and application of state-of-the-art, validated knowledge for defence and security purposes. S&T activities embrace scientific research, technology development, transition, application and field-testing, experimentation and a range of related scientific activities that include systems engineering, operational research and analysis, synthesis, integration and validation of knowledge derived through the scientific method.

In NATO, S&T is addressed using different business models, namely a collaborative business model where NATO provides a forum where NATO Nations and partner Nations elect to use their national resources to define, conduct and promote cooperative research and information exchange, and secondly an in-house delivery business model where S&T activities are conducted in a NATO dedicated executive body, having its own personnel, capabilities and infrastructure.

The mission of the NATO Science & Technology Organization (STO) is to help position the Nations' and NATO's S&T investments as a strategic enabler of the knowledge and technology advantage for the defence and security posture of NATO Nations and partner Nations, by conducting and promoting S&T activities that augment and leverage the capabilities and programmes of the Alliance, of the NATO Nations and the partner Nations, in support of NATO's objectives, and contributing to NATO's ability to enable and influence security and defence related capability development and threat mitigation in NATO Nations and partner Nations, in accordance with NATO policies.

The total spectrum of this collaborative effort is addressed by six Technical Panels who manage a wide range of scientific research activities, a Group specialising in modelling and simulation, plus a Committee dedicated to supporting the information management needs of the organization.

- AVT Applied Vehicle Technology Panel
- HFM Human Factors and Medicine Panel
- IST Information Systems Technology Panel
- NMSG NATO Modelling and Simulation Group
- SAS System Analysis and Studies Panel
- SCI Systems Concepts and Integration Panel
- SET Sensors and Electronics Technology Panel

These Panels and Group are the power-house of the collaborative model and are made up of national representatives as well as recognised world-class scientists, engineers and information specialists. In addition to providing critical technical oversight, they also provide a communication link to military users and other NATO bodies.

The scientific and technological work is carried out by Technical Teams, created under one or more of these eight bodies, for specific research activities which have a defined duration. These research activities can take a variety of forms, including Task Groups, Workshops, Symposia, Specialists' Meetings, Lecture Series and Technical Courses.

The content of this publication has been reproduced directly from material supplied by STO or the authors.

Published June 2020

Copyright © STO/NATO 2020
All Rights Reserved

ISBN 978-92-837-2253-3

Single copies of this publication or of a part of it may be made for individual use only by those organisations or individuals in NATO Nations defined by the limitation notice printed on the front cover. The approval of the STO Information Management Systems Branch is required for more than one copy to be made or an extract included in another publication. Requests to do so should be sent to the address on the back cover.

Table of Contents

	Page
List of Figures	**xii**
List of Tables	**xv**
List of Acronyms	**xviii**
Preface	**xxii**
Foreword	**xxiii**
Introduction	**xxiv**
Acknowledgements	**xxvi**
Postscript	**xxvii**
SAS-114 Membership List	**xxviii**
Executive Summary and Synthèse	**ES-1**
Part I: Organizational Aspects of Intelligence Production Management	**Part I-i**
Chapter 1 – Correcting Judgment Correctives in National Security Intelligence	**1-1**
1.1 Introduction	1-1
1.2 The IC's Corrective Approach	1-1
1.3 Critique of the Current Approach	1-2
1.3.1 The Incidental Approach to IC Innovation	1-2
1.3.2 Organizational Limitations	1-2
1.3.3 Conceptual Limitations	1-3
1.3.4 Correcting the IC's Current Corrective Approach	1-4
1.4 References	1-5
Chapter 2 – Mitigating Risk in the Analytic Workflow: A UK Perspective	**2-1**
2.1 Introduction	2-1
2.2 Defence Intelligence in the UK	2-1
2.2.1 Threat Assessment and Risk Assessment in DI	2-1
2.3 The Analytic Risk Framework	2-2
2.3.1 Intelligence Failure as a Risk	2-2
2.3.2 Assessing the Risk of Intelligence Failure	2-2
2.3.2.1 Cost to the Analyst	2-2
2.3.2.2 Cost to DI	2-2

2.3.2.3 Cost to the UK Government 2-2
2.3.2.4 Cost to Wider Society 2-3
2.3.3 Application to DI 2-3
2.4 Requirement 2-3
2.4.1 Outline 2-3
2.4.2 Mitigation 2-4
2.4.3 SAS-114 Contribution 2-4
2.5 Quantity of Information 2-5
2.5.1 Outline 2-5
2.5.2 Mitigation 2-6
2.5.3 SAS-114 Contribution 2-6
2.6 Quality of Information 2-6
2.6.1 Outline 2-6
2.6.2 Mitigation 2-7
2.6.3 SAS-114 Contribution 2-7
2.7 Judgment 2-8
2.7.1 Outline 2-8
2.7.2 Mitigation 2-9
2.7.3 SAS-114 Contribution 2-10
2.8 Output 2-11
2.8.1 Outline 2-11
2.8.2 Mitigation 2-12
2.8.3 SAS-114 Contribution 2-12
2.9 Summary 2-13
2.10 References 2-13

Chapter 3 – Devil's Advocacy Within Dutch Defence: Improving Intelligence Support to Decision Making 3-1

3.1 Introduction 3-1
3.2 Enter the Devil's Advocate: Concept and Methodology 3-1
3.2.1 Concept 3-2
3.2.2 Methodology 3-2
3.3 The Dutch Experience 3-3
3.3.1 Devil's Advocacy in Israel 3-3
3.3.2 Devil's Advocacy in the Netherlands 3-4
3.4 Devil's Advocacy on the Horizon? 3-7
3.5 Conclusion 3-8
3.6 References 3-9

Chapter 4 – Intelligence Professionals' Views on Analytic Standards and Organizational Compliance 4-1

4.1 Introduction 4-1
4.1.1 The Present Research 4-2
4.2 Method 4-3
4.2.1 Participants 4-3

4.2.2 Measures 4-3
4.2.2.1 ICD 203 Scales 4-3
4.2.2.2 Job Satisfaction 4-5
4.2.2.3 The Big Five Inventory – Conscientiousness Subscale 4-5
4.2.2.4 The Actively Open-Minded Thinking Scale 4-5
4.2.2.5 Organizational Commitment Scale 4-6
4.2.3 Procedure 4-6
4.3 Results 4-6
4.3.1 Scale Characteristics 4-6
4.3.2 Responses to ICD 203 Scales 4-8
4.3.3 ICD 203 Scale Correlates 4-9
4.4 Discussion 4-10
4.5 Conclusion 4-12
4.6 References 4-12

Chapter 5 – Introducing an Evidence-Based Approach to Analytical Tradecraft Training 5-1

5.1 Introduction 5-1
5.2 A Brief History of Our Analytical Thinking Training 5-1
5.3 An Evidence-Based Approach to Intelligence Analysis 5-3
5.3.1 How Should, and How Do, Analysts Structure Their Work? 5-4
5.3.2 What Strategies Should, and What Strategies Do, Analysts Use to Solve Problems? 5-6
5.3.3 Development and Validation of the Analysis Support Guide 5-6
5.3.4 Testing the Analysis Support Guide 5-7
5.3.5 Embedding Good Practice in Training and Assessment 5-7
5.4 Conclusion 5-8
5.5 Acknowledgements 5-9
5.6 References 5-9

Part II: Information Evaluation Under Uncertainty Part II-i

Chapter 6 – Applying Information Theory to Validate Commanders' Critical Information Requirements 6-1

6.1 Introduction 6-1
6.2 Background: The NATO Intelligence Community Dilemma 6-1
6.2.1 Establishing Command Information Priorities 6-2
6.3 The SAT Approach 6-2
6.4 The Indicators Validator™ SAT 6-4
6.4.1 Analysis of the IV SAT 6-5
6.5 Information Gain: A Principled Approach to Evaluating Indicator Usefulness 6-6
6.5.1 Computing Information Gain 6-6
6.6 Applying IV and Information Gain to the NATO Example 6-7

6.7 Discussion 6-8
6.8 References 6-9

Chapter 7 – Standards for Evaluating Source Reliability and Information Credibility in Intelligence Production 7-1

7.1 Introduction 7-1
7.2 Overview of Current Standards 7-1
7.2.1 Semantic Issues 7-2
7.2.2 Source Reliability Determinants 7-4
7.2.3 Information Credibility Determinants 7-5
7.3 Conceptualizing Information Quality 7-7
7.4 Alternative Approaches to Information Evaluation 7-10
7.5 References 7-13

Chapter 8 – A Reliability Game for Source Factors and Situational Awareness Experimentation 8-1

8.1 Introduction 8-1
8.2 Motivation 8-2
8.3 The Reliability Concept 8-3
8.4 The Reliability Game Overall Design 8-3
8.4.1 World Design 8-4
8.4.2 System Design 8-5
8.4.3 Content Design 8-8
8.4.4 Game Design Constraints 8-9
8.5 Game Evaluation 8-10
8.5.1 Experiment Set-Up 8-10
8.5.2 Feedback and Observations on the Game Design 8-12
8.5.3 Outcomes on Source Quality Rating 8-12
8.5.4 Outcomes on Confidence Rating 8-14
8.5.5 Outcomes on Card Positions 8-16
8.6 Conclusion 8-17
8.7 References 8-17

Chapter 9 – The Risk Game: Capturing Impact of Information Quality on Human Belief Assessment and Decision Making 9-1

9.1 Introduction 9-1
9.2 Game Design 9-2
9.2.1 Game-Play 9-3
9.2.2 Scenario-Based Game Design 9-4
9.2.2.1 Underlying Reasoning 9-4
9.2.2.2 Other Applications 9-6
9.2.3 Attributes 9-6
9.2.4 Sources of Information 9-6
9.2.5 Information Quality Dimensions 9-7

9.2.6 Levels of Information Quality 9-8
9.2.7 Information Cards and Fusion 9-10
9.2.8 Gathering of Belief States and Decision Making 9-11
9.3 Methodology Assessment 9-12
9.4 Exploratory Data Analysis 9-12
9.4.1 Analysis of Queries 9-13
9.4.1.1 Information Needs 9-13
9.4.1.2 Query Strategies 9-15
9.4.2 Belief States and Decision 9-16
9.4.3 The Effect of Information Quality 9-18
9.4.3.1 Falseness 9-18
9.4.3.2 Certainty and Precision 9-19
9.4.3.3 Relevance 9-20
9.4.4 Possible Biases 9-20
9.5 Conclusion 9-21
9.6 Acknowledgements 9-22
9.7 References 9-22

Part III: Intelligence and Risk Assessment Under Uncertainty **Part III-i**

Chapter 10 – Systematic Monitoring of Forecasting Skill in Strategic Intelligence **10-1**

10.1 Why Systematic Monitoring Matters 10-1
10.2 Geopolitical Forecasting Skill 10-2
10.3 Systematic Monitoring of Forecasting Skill in Canada 10-3
10.3.1 Initial Phase 10-3
10.3.2 Second Phase 10-4
10.3.3 Third Phase 10-7
10.4 Conclusion and Recommendations 10-10
10.5 References 10-12

Chapter 11 – Information Security Continuous Monitoring (ISCM) – Prioritizing Risks in Defensive Cyber Operations (DCO) **11-1**

11.1 Introduction to ISCM for DCO 11-1
11.1.1 The Risk Management Framework (RMF) vs. Security Authorization vs. ISCM 11-2
11.1.2 The Current US Government Approach for ISCM 11-5
11.1.2.1 ISCM for Non-High-Impact Systems: DHS CDM Program 11-5
11.1.2.2 ISCM for High-Impact Systems: Agency-Specific ISCM Program 11-5
11.2 ISCM Risk Scoring Methodology (Notional) 11-6
11.2.1 Compliance-Based vs. Performance-Based vs. Risk-Based 11-6
11.2.2 Dynamic-State Risk Posture 11-6

11.3 Top Ten Challenges: ISCM and Risk Scoring Methodology 11-8
11.3.1 Challenge #1 – Managing the Magnitude of the Attack Surface to be Monitored 11-8
11.3.2 Challenge #2 – Prioritizing the Asset and Data (Elements and Sets) to Be Monitored 11-8
11.3.3 Challenge #3 – Defining Monitoring Frequencies 11-10
11.3.4 Challenge #4 – Integrating Existing ISCM Capabilities 11-10
11.3.5 Challenge #5 – Prioritizing Security Automation Domains 11-11
11.3.6 Challenge #6 – Standardizing Data Analysis and Presentation Requirements 11-11
11.3.7 Challenge #7 – Defining Continuous Monitoring Dashboard Requirements 11-11
11.3.8 Challenge #8 – Defining ISCM Reporting Requirements 11-11
11.3.9 Challenge #9 – Establishing Performance Benchmarks 11-12
11.3.10 Challenge #10 – Collecting, Reusing, and Sharing of Data 11-12
11.4 Enhancing Intelligence Assessment by Means of an ISCM Framework 11-13
11.4.1 Is It Possible for the Intelligence Community to Benefit from a Methodology Similar to ISCM? 11-14
11.4.2 Benefits of Implementing a Continuous Monitoring Framework to Enhance Intelligence Assessments 11-14
11.5 References 11-15

Chapter 12 – Boosting Intelligence Analysts' Judgment Accuracy: What Works, What Fails? 12-1

12.1 Introduction 12-1
12.2 The Analysis of Competing Hypotheses Technique 12-2
12.3 The Present Research 12-3
12.4 Method 12-5
12.4.1 Participants 12-5
12.4.2 Design and Procedure 12-5
12.4.3 Materials 12-5
12.4.4 Coherentization and Coherence Weighting 12-7
12.4.5 Metrics 12-8
12.5 Results 12-9
12.5.1 Coherence of Probability Judgments 12-9
12.5.2 Accuracy of Probability Judgments 12-10
12.5.3 Information Usefulness 12-11
12.5.4 Recalibrating Probability Judgments 12-11
12.5.5 Aggregating Probability Judgments 12-12
12.6 Discussion 12-14
12.7 References 12-17

Chapter 13 – Intelligence and the Analysis of Narratives 13-1

13.1 Introduction 13-1
13.2 Critical Realism, Causality, and the Role of Language in Security and Intelligence 13-1

13.2.1 Critical Discourse Analysis 13-4
13.2.2 Securitization 13-5
13.3 Analysis by Contrasting Narratives 13-8
13.3.1 Possible Objections from the Field 13-10
13.4 Conclusion 13-13
13.5 References 13-13

Chapter 14 – Deductive Multi-Valued Logics for Practical Reasoning **14-1**
14.1 Reasoning Under Uncertainty 14-1
14.1.1 Issues with Induction 14-1
14.2 Deduction 14-3
14.3 Multi-Valued Logics 14-5
14.3.1 Tense Logic 14-7
14.3.2 Design of an Experiment with Tense Logic 14-9
14.3.3 Evaluation of Results on Tense Logic 14-13
14.4 Discussion with Suggestions for Software 14-14
14.5 References 14-16

Chapter 15 – When Tradecraft Won't Work: Describing the Bounds for Analytic Technique **15-1**
15.1 References 15-9

Part IV: Communicating Uncertainty in Intelligence Production **Part IV-i**

Chapter 16 – Issues of Uncertainty in Natural Language Communications **16-1**
16.1 Introduction 16-1
16.2 Overview of Uncertainty in Text 16-3
16.2.1 Uncertainty Within the Content 16-4
16.2.1.1 Imprecision and Vagueness 16-5
16.2.1.2 Ambiguity and Polysemy 16-6
16.2.1.3 Which Language? 16-6
16.2.2 Uncertainty About the Content 16-7
16.2.2.1 "Words of Estimative Probability" 16-8
16.2.2.2 Hedges and Other Evidential Markers 16-9
16.2.2.3 Passive Voice, Depersonalization, Time, etc. 16-11
16.3 Quantifying Evidentials for Credibility Weighting 16-12
16.3.1 A Brief Look at Studies Assigning Values to Words of Estimative Probability 16-13
16.3.2 Relative Weightings of Other Evidential Markers 16-15
16.3.3 A Few Examples 16-21
16.3.4 A Note on the Numbers 16-22

16.4 Summary 16-23
16.5 References 16-23

Chapter 17 – UK and US Policies for Communicating Probability in Intelligence Analysis: A Review 17-1
17.1 Introduction 17-1
17.2 Existing Policies for Communicating Probability in Intelligence Analysis 17-2
17.3 The IC's Commitment to Linguistic Probabilities 17-5
17.4 The IC Should Develop Evidence-Based Policies 17-6
17.5 Conclusion and the Way Forward 17-8
17.6 References 17-9

Chapter 18 – Variants of Vague Verbiage: Intelligence Community Methods for Communicating Probability 18-1
18.1 Introduction 18-1
18.2 Overview of Current Standards 18-2
18.2.1 National Security Intelligence Standards 18-2
18.2.1.1 NATO Standards 18-2
18.2.1.2 Canadian Standards 18-2
18.2.1.3 US Standards 18-3
18.2.1.4 UK Standards 18-7
18.2.1.5 Norwegian Standards 18-8
18.2.1.6 Dutch Standards 18-8
18.2.1.7 Danish Standards 18-9
18.2.2 Risk Management Standards 18-10
18.2.2.1 Canadian Risk Standards 18-10
18.2.2.2 US Risk Standards 18-11
18.2.2.3 UK Risk Standards 18-14
18.2.2.4 Dutch Risk Standards 18-15
18.2.3 Standards Used in Other Domains 18-17
18.3 Terminological Issues 18-20
18.4 Scoring Issues 18-23
18.5 References 18-24

Chapter 19 – How Intelligence Organizations Communicate Confidence (Unclearly) 19-1
19.1 Introduction 19-1
19.2 Overview of Current Standards 19-2
19.2.1 National Security Intelligence Standards 19-2
19.2.1.1 NATO Standards 19-2
19.2.1.2 Canadian Standards 19-2
19.2.1.3 US Standards 19-4
19.2.2 Standards Used in Other Domains 19-6
19.3 Terminological Issues 19-11
19.4 Convergence 19-12
19.5 Additional Confidence Determinants 19-13

19.6 Scale Structure 19-14
19.7 References 19-15

Chapter 20 – Communicating Analysis in the Digital Era and the Communication of Uncertainty in Commercial Open-Source Intelligence 20-1

20.1 The Changing Communications Environment 20-1
20.2 The Intelligence User Experience (UX) 20-1
20.3 Multimedia Competencies and the Intelligence Analyst 20-4
20.4 Wireframing an Intelligence Report 20-5
20.5 Structured Briefings and Face-to-Face Interaction 20-8
20.6 The Communication of Analysis and Uncertainty in Commercial Open-Source Intelligence Publications: The Case of *Jane's Intelligence Review* 20-8
20.7 The Visual Communication of Uncertainty 20-10
20.8 Conclusion 20-11
20.9 References 20-11

List of Figures

Figure		Page
Figure 4-1	Mean Ratings on ICD203-PVS	4-8
Figure 4-2	Mean Ratings on ICD203-OCS	4-9
Figure 4-3	Mean Difference in ICD 203 Scores	4-9
Figure 6-1	Hierarchy of Information Requirements	6-2
Figure 6-2	The Indicators Validator™ Model	6-4
Figure 7-1	Rogova's Ontology of Quality of Information Content	7-9
Figure 7-2	Rogova's Ontology of Quality of Information Sources	7-10
Figure 8-1:	MV Red Horizon Information and Its Track Before AIS Contact Loss as Displayed in the Scenario Map by the Red Line	8-4
Figure 8-2	Game Board on Which the Cards Need to be Positioned	8-7
Figure 8-3	Diagram of a Session of Reliability Game	8-7
Figure 8-4	Example of the Presentation of the Same Message in the Four Different Rounds	8-8
Figure 8-5	Flashcards – Vessel of Interest, Source Quality Levels and Confidence Levels in the Analysis	8-9
Figure 8-6	Example of a Picture Collected at the End of a Round	8-11
Figure 8-7	Players' Feedback Questionnaire Outcomes	8-12
Figure 8-8	Example of Source Quality Rating (Round 3) by Three Different Players	8-13
Figure 8-9	Source Quality Ratings by Card	8-14
Figure 8-10	Example of Confidence Rating by Different Players	8-15
Figure 8-11	Confidence Ratings by Hypothesis in the Different Rounds	8-15
Figure 9-1	A Player Assessing the Threat from an Unknown Track in the Vicinity of the Port, Based on Information Previously Queried	9-3
Figure 9-2	Board and Scenario for the Risk Game	9-5
Figure 9-3	Independent, Latent and Dependent Variables of the Risk Game	9-9
Figure 9-4	An Example of an Information Card Abstracting a Piece of Information About the Location of Vessel *A* Provided by the Tracker Processing the Radar Signal	9-10
Figure 9-5	Belief Assessment Form to be Filled in by the Player After Discovering Each Piece of Information	9-11
Figure 9-6	Players' Feedback After the Risk Game	9-12

Figure 9-7	Number of Queried Pieces of Information Based on the Final Decision Made	9-13
Figure 9-8	Ratio of Switches in Queried Information Between Vessel *A* and Vessel *B*	9-15
Figure 9-9	The Redundancy/Complementarity Values Grouped, Based on the Decision Made	9-16
Figure 9-10	Two Examples of Sequential Belief Assessments for Both Events *A* and *B* for Two Players	9-16
Figure 9-11	Final Belief State Before Decision	9-17
Figure 9-12	Decision and Relation to Final Belief State	9-18
Figure 9-13	Impact of False Information Ratio on the Final Belief State	9-19
Figure 9-14	Impact of Information Content Only on the Belief Change	9-19
Figure 9-15	Impact of Attribute on Belief Change from *A* to *B* or *B* to *A*	9-20
Figure 10-1	Calibration Curve and Recalibration Curve	10-5
Figure 10-2	Receiver-Operator Characteristic Curves by Unit	10-8
Figure 10-3	Model-Based Calibration Curves by Unit	10-9
Figure 11-1	ARL ISCM Risk Scoring Strategy	11-6
Figure 11-2	ISCM Risk Scoring User Interface	11-7
Figure 11-3	ISCM Risk Scoring Widgets	11-7
Figure 11-4	What to Continuously Monitor	11-9
Figure 12-1	Accuracy of Probability Judgments by Group Size and Aggregation Method	12-12
Figure 12-2	Probability of Improvement Achieved by Increasing Group Size by One Member	12-13
Figure 13-1	Aspects of Social Events Serve, Within a Certain Social Context, as Heuristic Artefacts to Mobilize Particular Perspectives That Enable New Social Events to Occur	13-9
Figure 13-2	An Example of ACN	13-10
Figure 14-1	Three Historical Maps of Canada's Political Boundaries, from 1667, 1763, and 1898	14-10
Figure 14-2	Choices Among Truth Values for Each Proposition may be Facilitated by an Easy and Colourful Arrangement of the Alternatives	14-11
Figure 15-1	Theory/Data Framework for Research and Analysis	15-3
Figure 16-1	Sentence-Level Uncertainty in Natural Language Communications	16-4
Figure 16-2	Probabilities Assigned by CIA Analysts to Various Hedges	16-13
Figure 16-3	Ranges of Percentages Assigned to Hedges by Analysts in Training	16-14

Figure 16-4	Weights Assigned to Expressions of Uncertainty Used in the Context of Medical Discussions Between Paediatricians and the Parents of Sick Children	16-14
Figure 16-5	Words of Estimative Probability	16-15
Figure 16-6	Ranking of *Unlikely* and *Likely* Modified by Booster *Very* and Downtoner *Somewhat* for a Proposition *p*	16-17
Figure 16-7	Some Modifications to Figure 16-6, Including Labels and Expressions for Complete Uncertainty	16-17
Figure 16-8	Bipolar Scale Based on Showing Point of Maximum Uncertainty in the Center	16-18
Figure 16-9	A Numerical Scale for Certainty *p* is Untrue (-1.0) to Certainty *p* is True (1.0), with the Point of Maximum Uncertainty Assigned the Value Zero	16-18
Figure 16-10	Negation of Words of Estimative Probability	16-19
Figure 16-11	Example of Relative Weightings of Various Verbs Expressing Uncertainty about Propositional Information	16-20
Figure 16-12	Examples of Mappings of Relative Rankings onto a Scale of Words of Estimative Probability and Percentages	16-20
Figure 18-1	CFINTCOM Likelihood	18-4
Figure 18-2	NIE 2007 Estimates of Likelihood	18-4
Figure 18-3	DIA Expressing Analytic Certainty	18-6
Figure 18-4	NIC Judgments of Likelihood	18-7
Figure 18-5	DISS Degree of Probability	18-9
Figure 18-6	DDIS Degrees of Probability	18-9
Figure 18-7	UK Cabinet Office NRR Risk of Terrorist and Other Malicious Attacks	18-14
Figure 18-8	UK Cabinet Office NRR Other Risks	18-15
Figure 19-1	DIA Identifying Confidence	19-5
Figure 19-2	DIA Expressing Analytic Certainty	19-6
Figure 19-3	IPCC Supplemental Qualitative Uncertainty Terms	19-8
Figure 19-4	IPCC AR4 Guidance Notes Quantitatively Defined Levels of Understanding	19-9
Figure 19-5	IPCC AR4 WGIII Qualitative Definition of Uncertainty	19-10
Figure 19-6	IPCC AR5 WGI Confidence	19-10
Figure 20-1	Characteristics of the Emerging Communications Environment	20-2
Figure 20-2	Wireframe of a Multimedia Report Created with Balsamiq Mockups	20-6
Figure 20-3	The Five Planes of the User Experience	20-7
Figure 20-4	Jane's Country Risk Scale	20-10

List of Tables

Table		Page
Table 2-1	Assessment of Analytic Risk Associated with Requirement	2-3
Table 2-2	Assessment of Analytic Risk Associated with Quantity of Information	2-5
Table 2-3	Assessment of Analytic Risk Associated with Quality of Information	2-6
Table 2-4	Assessment of Analytic Risk Associated with Judgment	2-8
Table 2-5	Assessment of Analytic Risk Associated with Output	2-11
Table 4-1	Means, Standard Deviations and Reliability Coefficients for All Scales	4-6
Table 4-2	Factor Loadings of the ICD 203 Scale – Self-Perspective Condition	4-7
Table 4-3	Correlates of ICD 203 Scales	4-10
Table 5-1	The Evolution of Analytic Tradecraft Training	5-2
Table 5-2	The Generic Analytic Workflow	5-5
Table 6-1	IV Matrix for the Scenario	6-7
Table 6-2	Information Gain Assessment for the Scenario	6-8
Table 7-1	NATO AJP 2.1 2016 Source Reliability and Information Credibility Scales	7-2
Table 7-2	NATO STANAG 2511 Source Reliability Scale	7-2
Table 7-3	NATO STANAG 2511 Information Credibility Scale	7-3
Table 8-1	Reliability Game State	8-5
Table 8-2	Reliability Game View	8-6
Table 8-3	Example of Reliability Game Messages	8-8
Table 8-4	Participant Demographics and Characteristics	8-11
Table 8-5	Example of Card Positions Collected for Each Player	8-16
Table 9-1	Sources Coverage and Expertise	9-7
Table 9-2	Eight Information Quality Levels and Corresponding Randomization	9-9
Table 9-3	Ratio of Vessels Queried by the Players	9-13
Table 9-4	Ratio of Attributes Queried by the Players	9-14
Table 9-5	Ratio of Attribute Categories Queried by the Players	9-14
Table 9-6	Ratio of Sources Queried by the Players	9-14

Table 10-1	Probability Terms in the Lexicon and Numeric Probability Equivalents	10-6
Table 11-1	RMF Steps	11-3
Table 11-2	Security Authorization Categories	11-4
Table 11-3	ISCM Steps	11-4
Table 12-1	Informational Features of Experimental Task	12-6
Table 14-1	The Truth Table for Material Implication in a Two-Valued Propositional Calculus	14-4
Table 14-2	The Truth Table for Material Implication in a Three-Valued Calculus, Under Kleene's Interpretation of a Third Truth Value	14-6
Table 14-3	The Truth Table for Material Implication in a Three-Valued Calculus, Under Łukasiewicz's Interpretation of a Third Truth Value	14-6
Table 14-4	The Truth Table for Weak Equivalence in a Three-Valued Calculus, Under Kleene's Interpretation of a Third Truth Value	14-6
Table 14-5	The Truth Table for Weak Equivalence in a Three-Valued Calculus, under Łukasiewicz's Interpretation of a Third Truth Value	14-7
Table 14-6	The Truth Table for 'Exclusive Or' (P V Q) in Prior's Six-Valued Calculus *Q*	14-8
Table 14-7	The Truth Table for a Two-Place Connective for Implication in Prior's Six-Valued Calculus *Q*	14-8
Table 14-8	The Truth Table for Conjunction 'and' in Prior's Six-Valued Calculus *Q*	14-9
Table 16-1	Final Scale with Seven Categories of Probability Expressions Plus Their Calculated Probability Points	16-17
Table 17-1	Standardized Lexicon Developed by ODNI	17-3
Table 17-2	Standardized Lexicon Developed by PHIA	17-3
Table 18-1	NATO AJP-2.1 2016 Probability Levels	18-2
Table 18-2	IAS MEA Probability Mapping Standard	18-2
Table 18-3	DIA Expressing Likelihood	18-5
Table 18-4	ICD 203 Analytic Standard	18-6
Table 18-5	DI Uncertainty Yardstick	18-7
Table 18-6	PHIA Probability Yardstick	18-8
Table 18-7	Etterretningsdoktrinen Confidence Levels	18-8
Table 18-8	DDIS Degrees of Probability	18-9
Table 18-9	CF Risk Assessment Matrix	18-10
Table 18-10	CF Probability Categories	18-10

Table 18-11	Joint Chiefs of Staff Intelligence/Probability Assessment	18-12
Table 18-12	Joint Staff J-5 Military Risk Assessment Matrix	18-12
Table 18-13	DOD Deliberate Risk Assessment Worksheet	18-13
Table 18-14	DOD Recommended Likelihood Criteria	18-13
Table 18-15	Netherlands Category Breakdown of Likelihood of Threats	18-16
Table 18-16	Netherlands Category Breakdown of Likelihood of Hazards	18-16
Table 18-17	USGCRP CSSR 5OD Likelihood	18-17
Table 18-18	IPCC TAR WGI Likelihood Statements	18-18
Table 18-19	IPCC TAR WGII Confidence Statements	18-18
Table 18-20	IPCC AR4 (WGI)/AR5 (WGII) Likelihood Terminology	18-18
Table 18-21	IPCC AR4 (WGI)/AR5 (WGII) Description of Likelihood	18-19
Table 18-22	IPCC AR4 (WGI) Special Report on Emissions Scenarios Likelihood	18-19
Table 18-23	IPCC AR5 (WGI, WGII) Likelihood Scale	18-19
Table 18-24	Netherlands Environmental Assessment Agency Verbal Information	18-20
Table 18-25	EFSA Probability Guidance	18-20
Table 19-1	NATO AJP-2.1 2016 Confidence Levels	19-2
Table 19-2	CFINTCOM Confidence	19-3
Table 19-3	Public Safety Canada Confidence Level	19-3
Table 19-4	NIE 2007 Confidence in Assessments	19-4
Table 19-5	JP 2-0 Confidence in Analytic Judgments	19-4
Table 19-6	NIC Confidence in the Sources Supporting Judgments	19-6
Table 19-7	USGCRP CSSR Confidence Level	19-7
Table 19-8	IPCC Scale for Assessing State of Knowledge	19-7
Table 19-9	IPCC TAR WGI Levels of Confidence	19-8
Table 19-10	IPCC TAR WGII Qualitative Assessment of the State of Knowledge	19-9
Table 19-11	IPCC AR4 Guidance Notes Quantitatively Calibrated Levels of Confidence	19-9
Table 19-12	IPCC AR4/AR5 Description of Confidence	19-10
Table 19-13	IPCC AR5 Confidence	19-10
Table 19-14	WADA Levels of Confidence	19-11
Table 20-1	Application of Usability to Traditional Analytic Products	20-4
Table 20-2	Expressions of Probability in *Jane's Intelligence Review* According to the US ICD-203 of 2 January 2015	20-9

List of Acronyms

ABI	Activity-Based Intelligence
ACE	Aggregative Contingent Estimation
ACH	Analysis of Competing Hypotheses
ACN	Analysis by Contrasting Narratives
ACS	Affective Commitment Scale
ACT	Allied Command Transform
AD	Allocation Disagreement
ADS	Analytic Decision Science
AIS	Automatic Identification System
AO	Authorizing Official
AOR	Area of Responsibility
AOT	Actively Open-minded Thinking
AR4	Fourth Assessment Report
AR5	Fifth Assessment Report
ARL	Army Research Laboratory
ASG	Analysis Support Guide
BA	Belief Assessment
BDP	Big Data Platform
BFI	Big Five Inventory
CAF	Canadian Armed Forces
CAP	Coherent Approximation Principle
CCIR	Commander's Critical Information Requirement
CCS	Continuance Commitment Scale
CDA	Critical Discourse Analysis
CDM	Continuous Diagnostics and Mitigation
CED	Captured Enemy Documents
CELLEX	Cell Phone Exploitation
CFINTCOM	Canadian Forces Intelligence Command
CIA	Central Intelligence Agency
CMRE	Centre for Maritime Research and Experimentation
CSO	Collaboration Support Office
CSSR	Climate Science Special Report
CT	Counter Terrorism
DCO	Defensive Cyber Operations
DDIS	Danish Defence Intelligence Service
DHS	Department of Homeland Security
DI	Defence Intelligence
DISS	Defence Intelligence and Security Service
DNI	Director of National Intelligence
DNS	Domain Name Server
DOD	Department of Defense
DOMEX	Document Exploitation
DRDC	Defence Research and Development Canada
EEFI	Essential Elements of Friendly Information
EEI	Essential Enemy Information

EEZ	Exclusive Economic Zone
EFP	Enhanced Forward Presence
EFSA	European Food Safety Authority
ERM	Enterprise Risk Management
EU INTCEN	European Union Intelligence and Situation Centre
FFIR	Friendly Force Information Requirement
FISMA	Federal Information Security Management Act
G/P/S	Gameplay/Purpose/Scope
GEOINT	Geospatial Intelligence
GM	Game Mechanics
HUMINT	Human Intelligence
IARPA	Intelligence Advanced Research Projects Activity
IAS	Intelligence Assessment Secretariat
IC	Intelligence Community
ICA	Intention, Capability, Activity
ICD	Intelligence Community Directive
ICD203-OCS	Intelligence Community Directive 203 Organizational Compliance Scale
ICD203-PVS	Intelligence Community Directive 203 Professional Values Scale
IDF	Israeli Defence Force
IDS/IPS	Intrusion Detection/Prevention Systems
IM	Incoherence Metric
IMINT	Imagery Intelligence
IPCC	Intergovernmental Panel on Climate Change
IQ	Information Quality
IQD	Information Quality Dimensions
IR	Information Requirement
IR	International Relations
IRTPA	Intelligence Reform and Terrorism Prevention Act
IS	Information System
ISCM	Information Security Continuous Monitoring
IT	Information Technology
IV	Indicators Validator™
JDM	Judgement and Decision Making
JIC	Joint Intelligence Committee
JIR	Jane's Intelligence Review
JP	Joint Publication
KE	Knowledge Elicitation
LINOP	Linear Opinion Pool
MACE	Method for Assessing the Credibility of Evidence
MAE	Mean Absolute Error
ME	Mean Error
MEA	Middle East and Africa
MECE	Mutually Exclusive and Collectively Exhaustive
MEDEX	Media Exploitation
MPA	Marine Protected Area

NATO	North Atlantic Treaty Organization
NCM	Non Commissioned Member
NCS	Normative Commitment Scale
NIC	National Intelligence Council
NIE	National Intelligence Estimate
NIST	National Institute of Standards and Technology
NRR	National Risk Register
NSA	National Security Agency
OA	Ongoing Authorization
ODNI	Office of the Director of National Intelligence
OMB	Office of Management and Budget
ONR	Office of Naval Research
OoW	Officer of the Watch
OSINT	Open Source Intelligence
PCTEG	Policy Counter Terrorism Evaluation Group
PHIA	Professional Head of Intelligence Assessment
PIR	Priority Intelligence Requirement
PIT	Platform Information Technology
PKI	Public Key Infrastructure
PoI	Piece of Information
QD	Quantity Disagreement
RMF	Risk Management Framework
RMW	Risk Management Widget
ROC	Receiver-operator Characteristic
RTG	Research Task Group
SA	Situational Assessment
SA	Situational Awareness
SAR	Search and Rescue
SAT	Structured Analytic Technique
SAW	Situational Awareness
SD	Standard Deviation
SDLC	System Development Life Cycle
SIEM	Security Information and Event Management
SIGINT	Signals Intelligence
SME	Subject Matter Expert
SP	Special Publication
SSN	Sum of the (Signed) Non-additivity
STANAG	Standardization Agreement
TAR	Third Assessment Report
TDQM	Total Data Quality Management
TTX	Table Top Exercise
UK	United Kingdom
US	United States
USGCRP	US Global Change Research Program
UX	User Experience
VoIP	Voice over Internet Protocol

W&P	Weighting and Prioritizing
WADA	World Anti-Doping Agency
WEP	Word of Estimative Probability
WGI	Working Group I
WGII	Working Group II
WMD	Weapons of Mass Destruction
WYSIATI	What-You-See-Is-All-There-Is

Preface

SAS-114 was preceded by the Exploratory Team SAS-ET-CR, which met once at NATO's Collaboration Support Office (CSO) in December, 2014. The original idea was to focus on a review of standards for communicating uncertainty and risk. Following recommendations from Great Britain, the scope was broadened to include not only uncertainty communication but also how assessments are made under conditions of uncertainty. The SAS-114 Research Task Group (RTG) kicked off at CSO a year later with an initial team from Canada, Denmark, Great Britain, The Netherlands, and USA as well as from NATO's Centre for Maritime Research and Experimentation. Over time, it expanded its membership to include Germany, Norway, Spain, and Sweden. Within the first year, it also became apparent that SAS-114 was mainly focused on the domain of intelligence analysis. The emphasis on intelligence was formally captured in a mid-cycle renaming of the activity. SAS-114 picked up many new members from the intelligence community, and the team's composition became truly diverse, including a mix of scientists and intelligence professionals. Each meeting was structured like a small conference meant to exchange ideas, new findings and do something rarely done: give scientists and practitioners a bi-annual space of a few days to discuss challenges in intelligence analysis and to hear about cutting-edge research that could be used to improve intelligence and its communication to decision-makers. As a result, SAS-114 also benefited from a large and healthy dose of invited speakers from both the scientific and intelligence communities. A representative example was captured in the Meeting Proceedings, *Communicating Uncertainty, Assessing Information Quality and Risk, and Using Structured Techniques in Intelligence Analysis* (doi: 10.14339/STO-MP-SAS-114), which outlined a workshop hosted by Arne Biering at Kastellet in Copenhagen. The meeting structure of SAS-114, uncharacteristic of RTG meetings, was meant to spur frank and open dialogue and provide opportunities for collaborations to form and develop. The core team did not set out to design experiments that all members would contribute to. Rather, smaller clusters of collaborative effort formed where mutual interest was strong and each participating member had value to contribute. Many of the chapters in this report outline the culmination of such collaborative efforts. Some of those team efforts are still ongoing and not all of them have matured to the point where they can be summarized in this report. If SAS-114 had produced little over the last three years, this might rightly be construed as "unfinished business," yet by any reasonable standard, SAS-114 has been highly productive and the ongoing collaborations are more aptly construed as a clear sign of the team's sustained collaborative strength and generative potential, which will extend far beyond its predetermined years, perhaps even as one or more future NATO activities.

That SAS-114 proved to be an experiment in open dialogue and self-forming research collaborations is well reflected in this final report. In it, readers will find no consensus document co-signed by its members, but rather a diverse collection of structured analyses, research findings, professional insights, and thoughtpieces that bear on the key foci of SAS-114. As Editor, I have occasionally challenged authors on substantive points, but only to further sharpen arguments, never to force a common viewpoint. The twenty chapters in this report cover wide territory and are organized into four parts: (a) organizational aspects of intelligence production management, (b) information evaluation under uncertainty, (c) intelligence and risk assessment under uncertainty, and (d) communicating uncertainty in intelligence production. Whereas the last part squarely addresses the original aim of SAS-114, tracing back to the Exploratory Team, the first three parts highlight how the activity has developed since those early beginnings. This report is, in effect, an edited book, and readers are invited to approach it as such, reading those chapters that pique their interest while feeling no guilt in putting aside those that do not. Some of the chapters are reprints or adaptations of papers that have been published in peer-reviewed outlets, while others have been written specifically for this report. Among the latter, and attesting to the quality of the contributions, some have already found their way in adapted form into peer-reviewed publications. The hope of the SAS-114 team is that you find practical and intellectual value in this collection of papers.

David Mandel, March 19, 2019

Foreword

Commanders and policy-makers need quality information to make appropriate decisions. When dealing with their own forces, getting the right information at the right level and at the right time, while not trivial, can be achieved through management excellence. Risks can then be properly measured and accounted for. No amount of management excellence, however, can deliver decision-quality information about a competent adversary without injecting a significant amount of uncertainty.

Most of that uncertainty comes from not having access to the information first-hand and having to extrapolate from incomplete or proxy measurements – a situation that would be familiar to analysts in other walks of life, be it market research, operations research or financial analysis. Some of that uncertainty, however, comes from the adversary using active deception and attempting to turn our own biases against us to mask intentions and capabilities. To adapt the descriptor we use to depict adversary courses of action: if the first generator of uncertainty is the most likely, the second one is the most dangerous. Together, they provide two distinct but related challenges to the intelligence analyst: How to arrive at proper assessments under these conditions and How to properly communicate this uncertainty to decision-makers.

While intuitively coherent processes for promoting accuracy of intelligence assessments and communicating uncertainties have been in use for some time in most intelligence organizations, research presented in this report suggest that some do not stand up to the scrutiny of the scientific method. The recurring theme emanating from this collection of papers is that, as we continue to build on the growing understanding between researchers and practitioners, the evolving science of judgement and decision-making can help in the development of an evidence-based approach to intelligence analytic tradecraft.

A symbiotic relationship between science and warfare is not new. From the first cave dwellers experimenting with the size, shape and material to use for a stick to defend the family against assailants, to the development of the stealth aircraft, the link between research, development and defence practitioners has been strong in the operational functions of Act, Shield and Sense. The Command function, including its Intelligence subset, has proven more resistant to help from the scientific community. Culture, the difficulty to get scientists cleared at the appropriate level and the lack of appeal of publishing research so classified that it cannot be peer reviewed are some of the factors that have contributed to this distance.

The rich collection of thought pieces, professional insights and research findings in this Technical Report—the product of scientists and practitioners deliberately getting together to discuss challenges in intelligence analysis—is a sign that we are finally breaching that divide. Both tribes are sure to gain from this collaborative approach, but the biggest winner will undoubtedly be consumers of intelligence: commanders, policy-makers and the people they serve.

MGen (Ret'd) Christian Rousseau

Director of Canada's Integrated Terrorism Assessment Centre; former Canadian Chief of Defence Intelligence and founding Commander of the Canadian Forces Intelligence Command.

Introduction

Military and civilian intelligence organizations are routinely called on to support commanders and policymakers, whose decisions affect national and international security. Among other features, such as timeliness and relevance, intelligence organizations are meant to produce assessments that are supported by rigorous analysis, that are accurate, and that are communicated clearly to decision-makers. Uncertainty poses a key challenge to both the assessment and communication functions of intelligence. For instance, the quality of information that analysts receive is often uncertain as are the conceptual models on which they rely. In short, most analysis is human judgment made under conditions of uncertainty. Decision-makers may wish to fully eliminate uncertainty, but intelligence organizations must strive to communicate lingering uncertainties about events (probabilities) and about their assessments (confidence) as coherently and clearly as possible to avoid miscommunication.

The SAS-114 Research Task Group addressed these dual challenges by examining (a) existing and novel methods for promoting the accuracy of intelligence assessments under uncertainty and (b) standards for communicating uncertainties in such assessments. This report, which outlines the research and analysis completed by SAS-114, is organized into the following four parts:

Part I (chapters 1-5) examines organizational aspects of intelligence production management. Chapter 1 outlines how current intelligence training, as formulated by thought leaders with limited scientific expertise, fails to address the subjectivity inherent in uncertainty communication or encourage self-critical cognition on the part of analysts. Chapter 2 presents a framework for uncertainty evaluation meant to maximize value to decision makers and reduce the risk of intelligence failures, based on the experience of UK Defence Intelligence. Chapter 3 describes the Dutch Defence Intelligence and Security Service's use of Devil's Advocacy to improve analytic products. Chapter 4 presents research on the extent to which Canadian intelligence practitioners view themselves and their organizations as compliant with standards for analytic rigor mandated under US Intelligence Community Directive 203. In Chapter 5, members of the UK Analytical Tradecraft Training Team discuss how academic collaboration and internal research facilitate the implementation of evidence-based tradecraft within their organization.

Part II (chapters 6-9) of this report focuses on information evaluation under uncertainty. Chapter 6 presents a novel approach to establishing intelligence collection priorities based on expected information value. Chapter 7 critically examines current standards for evaluating source reliability and information credibility, and highlights avenues for future research. Next, chapter 8 introduces the Reliability Game as a gaming approach to measuring the impact of source factors on human situational awareness. Chapter 9 follows with a discussion of the Risk Game, a methodology to assess how experts process heterogeneous information, consider information quality, and form beliefs about concurrent events.

Part III (chapters 10-15) examines intelligence and risk assessment under uncertainty. Chapter 10 discusses the importance of systematically monitoring geopolitical forecasting skill and outlines empirical methods for doing so. Chapter 11 focuses on the challenges of information security continuous monitoring (ISCM) in defensive cyber operations, and discusses the application of an ISCM framework to improve intelligence assessments. Chapter 12 presents experimental research on the effectiveness of Analysis of Competing Hypotheses, as well as post-analytic recalibration and aggregation methods, as means of improving analysts' judgment accuracy. Chapter 13 introduces the theory of critical realism, along with the theoretical components of critical discourse analysis and securitization theory, which together provide a framework for a novel analytic methodology: analysis by contrasting narratives. Chapter 14 follows with a transparent method of combining analytic judgments in the form of truth tables for 3-valued and 6-valued logics. Chapter 15 concludes with a classification system that assists in mapping analytic techniques to specific intelligence problems.

Part IV (chapters 16-20) of this report discusses the communication of uncertainty in intelligence production, in line with the original aims of SAS-114. Chapter 16 examines how the uncertainty inherent in natural language affects reporting quality, and presents a methodology for identifying, evaluating, and weighting the evidentiality of textual information. Chapter 17 provides a critical review of US and UK policies for communicating probability in intelligence analysis. Chapter 18 presents an annotated collection of estimative probability standards gathered by members and affiliates of SAS-114. Similarly, Chapter 19 presents SAS-114's collection of standards used to assess and communicate analytic confidence. Chapter 20 concludes the report with a discussion of communication in the digital era, with a particular focus on uncertainty communication in commercial open source intelligence.

The twenty chapters in this report thus cover wide conceptual ground. The hope of the SAS-114 team is that the reader will find this collection both intellectually stimulating and of practical use.

Acknowledgements

SAS-114 wishes to thank Arne Biering (DNK) and Bob de Graaf (NLD) for their contributions to the early phase of this activity. Both Arne and Bob played seminal roles in shaping the intelligence-focused direction of SAS-114, and sadly, after Arne's retirement, SAS-114 lost DNK as a member nation. SAS-114 also thanks the many invited speakers who played a vital role in energizing and informing discussions at our biannual meetings. Among them, David Budescu (USA) deserves special thanks for presenting informative research at no less than three separate SAS-114 meetings and for co-designing research with the SAS-114 team that is still ongoing. The SAS-114 team is grateful to Army Research Laboratory (USA), the Centre for Maritime Research and Experimentation (NATO), the Collaboration Support Office (NATO), the Danish Defence Intelligence Service (DNK), Middlesex University (GBR) and Rey Juan Carlos University (ESP) for their generous hospitality and support in hosting our bi-annual meetings and workshops. We thank Tonya Hendriks (CAN) and William Kozey (CAN) for their assistance with preparing this report and the earlier meeting proceedings, and we extend a special thanks to Daniel Irwin (CAN), who with painstaking care, assisted the SAS-114 Chair in bringing this Technical Report to fruition. We thank the NATO CSO editing team at NIVA Inc. for their diligence in copy editing this report and managing it through the editorial process. We are especially indebted to LTC Timothy Povich, Rina Tahar, and Jeroen Groenevelt at the Collaboration Support Office for their guidance and assistance over the past several years. Finally, we thank the entire SAS Panel and the Office of the Chief Scientist for providing constructive feedback on an earlier draft of this report at the 2019 Fall Panel Business Meeting in Ottawa, Canada.

Postscript

The lag time between the submission of a manuscript and its eventual publication is something that most authors, with varying degrees of patience, learn to accept over time. As it is most often the case, this interval permitted the final report to be carefully honed and further improved. However, in this particular case, the interval has also afforded me the opportunity to convey here that the SAS-114 RTG was recently awarded the SAS Panel Excellence Award in recognition of the outstanding quality of the research conducted by our team. On behalf of the entire SAS-114 team, I thank the SAS Panel Awards Committee for recognizing our efforts in this way. Such news is a fine way to bring this activity to a formal close as we look towards the future and continue to explore ways of applying what we have learned.

David Mandel, May 13, 2020

SAS-114 Membership List

CO-CHAIRS

Dr. David R. MANDEL (Chair)*
Defence Research and Development Canada
CANADA
Email: david.mandel@drdc-rddc.gc.ca

MEMBERS

Dr. Rubén ARCOS*
Rey Juan Carlos University
SPAIN
Email: ruben.arcos@urjc.es

Dr. 2LT Jonas CLAUSEN MORK
Swedish Defence Research Agency
SWEDEN
Email: jonas.clausen.mork@foi.se

Dr. Alexander CLAVER*
Ministry of Defence
NETHERLANDS
Email: a.claver@mindef.nl

Ms. Francesca DE ROSA*
Centre for Maritime Research and Experimentation
ITALY
Email: francesca.derosa@cmre.nato.int

Dr. Peter DE WERD*
Netherlands Defence Academy
NETHERLANDS
Email: pg.d.werd@mindef.nl

Dr. Mandeep K. DHAMI*
Middlesex University
UNITED KINGDOM
Email: m.dhami@mdx.ac.uk

Ms. Tonya L. HENDRIKS*
Defence Research and Development Canada
CANADA
Email: tonya.hendriks@drdc-rddc.gc.ca

Mr. Daniel IRWIN*
Defence Research and Development Canada
CANADA
Email: dan.irwin@drdc-rddc.gc.ca

Cdr. Bjørn M. ISAKSEN
Royal Norwegian Navy
NORWAY
Email: b.isaksen@cranfield.ac.uk

Dr. Anne-Laure JOUSSELME*
Centre for Maritime Research and Experimentation
ITALY
Email: anne-laure.jousselme@cmre.nato.int

Dr./Lt Col James E. KAJDASZ*
USAF Academy, United States Air Force
UNITED STATES
Email: jimkaj@aol.com

Dr. Jonathan D. NELSON*
University of Surrey
UNITED KINGDOM
Email: jonathan.d.nelson@gmail.com

Dr. Keith K. NIALL*
Defence Research and Development Canada
CANADA
Email: keith.niall@drdc-rddc.gc.ca

Mr. Mark A. C. TIMMS*
Department of National Defence
CANADA
Email: mark.timms@forces.gc.ca

* Contributing Author

Dr. Kellyn REIN*
Fraunhofer FKIE
GERMANY
Email: kellyn.rein@fkie.fraunhofer.de

Ms. Ramine SHAW
Department of National Defence
CANADA
Email: ramine.shaw@forces.gc.ca

Dr. Huibert M.VAN DE MEEBERG*
Ministry of Defence
NETHERLANDS
Email: hm.vd.meeberg@mindef.nl

Dr. Raul VICEN
Centre for Maritime Research and Experimentation
ITALY
Email: raul.vicen@cmre.nato.int

Mr. Greg WEAVER*
Army Research Laboratory
UNITED STATES
Email: greg.a.weaver3.civ@mail.mil

ADDITIONAL CONTRIBUTORS

DI Futures and Analytical Methods*
Ministry of Defence
UNITED KINGDOM
Email : Email unavailable

Dr. Alessandro DE GLORIA*
University of Genoa
ITALY
Email: adg@elios.unige.it

Dr. Christopher W. KARVETSKI*
KaDSCi LLC
UNITED STATES
Email: ckarvetski@gmail.com

Mr. Jonathan LOCKE*
Centre for Maritime Research and Experimentation
ITALY
Email: locke@cmre.nato.int

Mr. Akhilomen O. ONIHA*
Army Research Laboratory
UNITED STATES
Email: akhilomen.o.oniha.civ@mail.mil

Dr. Giuliana PALLOTTA*
Centre for Maritime Research and Experimentation
ITALY
Email: giuliana.pallotta@cmre.nato.int

Dr. Philip E. TETLOCK*
University of Pennsylvania
UNITED STATES
Email: tetlock@wharton.upenn.edu

UK Analytical Tradecraft Training Team*
UNITED KINGDOM
Email: unavailable

* Contributing Author



Assessment and Communication of Uncertainty in Intelligence to Support Decision Making

(STO-TR-SAS-114)

Executive Summary

Military and civilian intelligence organizations are routinely called on to support commanders and policymakers, whose decisions affect national and international security. Among other features, such as timeliness and relevance, intelligence organizations are meant to produce assessments that are supported by rigorous analysis, that are accurate, and that are communicated clearly to decision makers. Uncertainty poses a key challenge to both the assessment and communication functions of intelligence. For instance, the quality of information that analysts receive is often uncertain as are the conceptual models on which they rely. In short, most analysis is human judgement made under conditions of uncertainty. Decision makers may wish to fully eliminate uncertainty, but intelligence organizations must strive to communicate lingering uncertainties about events (probabilities) and about their assessments (confidence) as coherently and clearly as possible to avoid miscommunication.

The SAS-114 Research Task Group addressed these dual challenges by examining:

a) Existing and novel methods for promoting the accuracy of intelligence assessments under uncertainty; and

b) Standards for communicating uncertainties in such assessments.

This report, which outlines the research and analysis completed by SAS-114, is organized into four parts:

a) Part I (Chapters 1 – 5) examines organizational aspects of intelligence production management;

b) Part II (Chapters 6 – 9) examines information evaluation under uncertainty;

c) Part III (Chapters 10 – 15) examines intelligence and risk assessment under uncertainty; and

d) Part IV (Chapters 16 – 20) examines current methods of communicating uncertainty in intelligence production.

A central theme of Part I is that intelligence organizations need to be proactive in leveraging the science of judgement and decision making. Part I further illustrates many ways in which intelligence organizations in Allied countries are attempting to develop a more evidence-based approach to analytic tradecraft and intelligence oversight. Part II critically examines current intelligence methods for evaluating information usefulness and quality, and it proposes alternative methods. Part II also introduces research methods for testing how analysts evaluate information quality in uncertain environments. Part III describes methods for monitoring the accuracy of intelligence forecasts and for monitoring defensive cyber risk. Part III also devotes significant attention to alternative methods for supporting intelligence analysis, including through support to the analyst but also through post-analytic methods drawn from decision science. Part IV zeroes in on the communication of uncertainty in natural language and in the intelligence domain. Several chapters offer critical analyses of current intelligence (and other professional) standards for communicating probabilities and confidence levels to decision makers.

Despite the diversity of topics and investigative approaches covered in this report, several chapters converge on some key conclusions. First, existing methods for communicating uncertainties about information quality,

event occurrence, and assessment accuracy are flawed in multiple respects that should prompt intelligence communities under NATO to pay closer attention to the relevant science. Specifically, we recommend that intelligence organizations consider using numeric probabilities instead of the vague verbal expressions of uncertainty currently in use. Second, we recommend that intelligence organizations test the effectiveness of analytic tradecraft methods in experiments that meet scientific standards, and that they consider alternative methods that have a stronger basis in scientific theory. This is vital because, as some of our research has shown, existing methods may not only fail to improve analytic rigour, they may in fact weaken the quality of analysts' assessments. Finally, we recommend that intelligence organizations adopt proactive systems of self-monitoring, tracking among other things, the accuracy of the forecasts they provide to decision makers.

Évaluation et communication de l'incertitude dans le renseignement en vue de faciliter la prise de décision

(STO-TR-SAS-114)

Synthèse

Les organisations de renseignement militaires et civiles sont régulièrement sollicitées pour éclairer les commandants et décideurs, dont les décisions influent sur la sécurité nationale et internationale. Parmi d'autres caractéristiques, telles que la rapidité d'information et la pertinence, les organisations de renseignement sont censées produire des évaluations précises, étayées par une analyse rigoureuse, qui sont communiquées clairement aux décideurs. L'incertitude pose un problème essentiel aux deux fonctions du renseignement que sont l'évaluation et la communication. Par exemple, la qualité des informations reçues par les analystes est souvent incertaine, tout comme les modèles conceptuels sur lesquels elles se basent. En bref, la majorité de l'analyse consiste en un jugement humain réalisé dans des conditions d'incertitude. Les décideurs peuvent souhaiter éliminer l'incertitude, mais les organisations de renseignement doivent s'efforcer de communiquer les incertitudes persistantes sur les événements (probabilités) et leur évaluation (confiance), de façon aussi cohérente et claire que possible pour éviter une erreur de communication.

Le groupe de recherche SAS-114 s'est penché sur ces deux défis en examinant :

a) Les méthodes existantes et inédites pour promouvoir l'exactitude des évaluations de renseignement dans le cadre de l'incertitude ; et

b) Les normes de communication des incertitudes dans ces évaluations.

Le présent rapport, qui décrit les recherches et l'analyse menées par le SAS-114, est organisé en quatre parties :

a) La partie I (chapitres 1 à 5) étudie les aspects organisationnels de gestion de la production du renseignement ;

b) La partie II (chapitres 6 à 9) examine l'évaluation des informations dans le cadre de l'incertitude ;

c) La partie III (chapitres 10 à 15) s'intéresse à l'évaluation du renseignement et du risque dans le cadre de l'incertitude ; et

d) La partie IV (chapitres 16 à 20) traite des méthodes actuelles de communication de l'incertitude dans la production du renseignement.

L'un des thèmes centraux de la partie I est que les organisations de renseignement doivent être proactives dans l'exploitation de la science du jugement et de la prise de décisions. La partie I illustre de nombreuses manières dont les organisations de renseignement des pays alliés tentent d'élaborer une approche de savoir-faire analytique et de supervision du renseignement reposant davantage sur des éléments tangibles. La partie II examine les méthodes actuelles du renseignement servant à évaluer l'utilité et la qualité des informations, et propose des méthodes alternatives. La partie II présente également des méthodes de recherche pour tester la façon dont les analystes évaluent la qualité des informations dans les environnements incertains. La partie III décrit des méthodes de suivi de l'exactitude des prévisions du renseignement et de suivi du risque de cyberdéfense. La partie III accorde également une grande attention aux méthodes

alternatives soutenant l'analyse du renseignement, notamment par le soutien apporté à l'analyste, mais aussi par des méthodes post-analytiques tirées de la science de la décision. La partie IV cerne la question de la communication de l'incertitude dans le langage naturel et dans le domaine du renseignement. Plusieurs chapitres offrent une analyse critique des normes actuelles du renseignement (et d'autres normes professionnelles) pour communiquer aux décideurs les probabilités et le niveau de confiance.

Malgré la diversité des sujets et des démarches d'investigation traités dans ce rapport, plusieurs chapitres convergent vers des conclusions essentielles. Tout d'abord, les méthodes existantes de communication des incertitudes sur la qualité des informations, la survenue d'événements et l'exactitude de l'évaluation sont biaisées à plusieurs égards, ce qui devrait inciter les communautés du renseignement de l'OTAN à s'intéresser de plus près à la science correspondante. Nous recommandons en particulier que les organisations de renseignement envisagent d'utiliser des probabilités numériques au lieu des expressions verbales floues actuellement utilisées pour décrire l'incertitude. Ensuite, nous recommandons que les organisations de renseignement testent l'efficacité de leurs méthodes d'analyse dans des expériences répondant à des normes scientifiques et qu'elles envisagent d'autres méthodes plus étayées sur le plan de la théorie scientifique. Il est vital de le faire parce que, comme l'ont montré certaines de nos recherches, les méthodes existantes peuvent non seulement ne pas améliorer la rigueur analytique, mais amoindrir la qualité de l'évaluation des analystes. Enfin, nous recommandons que les organisations de renseignement adoptent des systèmes proactifs d'autosurveillance, en suivant notamment l'exactitude des prévisions qu'elles fournissent aux décideurs.

Part I: ORGANIZATIONAL ASPECTS OF INTELLIGENCE PRODUCTION MANAGEMENT



Chapter 1 – CORRECTING JUDGMENT CORRECTIVES IN NATIONAL SECURITY INTELLIGENCE[1,2]

David R. Mandel
Defence Research and Development Canada
CANADA

Philip E. Tetlock
University of Pennsylvania
UNITED STATES

1.1 INTRODUCTION

Intelligence organizations in government play a vital role in informing the upper echelons of policymaking, the leaders of nations and their staff who are vested with the responsibility of protecting national security and promoting national interests. Within a given nation, the collective of intelligence organizations – euphemistically known as the Intelligence Community or, simply, the IC – therefore has an epistemic mandate to deliver timely, relevant, and accurate information to decision makers who operate under time and accountability pressures, the fog of uncertainty, and with foreknowledge that their decisions may alter the course of history.

How then has the IC sought to guarantee for policymakers and the public that they are doing their best to meet their epistemic mandate, given that the vast majority of substantive intelligence relies on human judgments made under conditions of deep uncertainty [3]? Do the IC's tactics to ensure judgment quality rest on sound strategy properly informed by key concepts, methods and findings from judgment and decision science, the field that speaks directly to the challenges the IC faces? To the latter question, we believe the answer is – No. Yet we also remain optimistic that the IC could substantially improve the quality of its judgments if it took appropriate steps to correct its current corrective strategy – steps that we lay out as a set of IC policy prescriptions.

1.2 THE IC'S CORRECTIVE APPROACH

The IC is well aware both that its primary analytic product is judgment to support decision making and that human judgment is prone to bias and error. Sherman Kent, an historian recruited to the fledgling IC during World War II and now widely regarded as the founder of modern intelligence analysis, was keenly concerned about the threats that confirmation bias and groupthink posed to epistemic integrity [4]. Richards Heuer Jr. went further, documenting in *Psychology of Intelligence Analysis* [5] how cognitive biases, much of which were revealed in the heuristics and biases research program inspired by Daniel Kahneman and Amos Tversky [6], could skew intelligence judgments and raise the risk of intelligence failure.

Heuer and others improvised simple, back-of-the-napkin, judgment-support methods that analysts could self-apply to debias their judgments and consequently improve their accuracy. The methods, which came to be known as structured analytic techniques or SATs, have proliferated [7] and continue to represent the IC's main tactical approach to combatting judgment error. In the US, the Intelligence Reform and Terrorism Prevention Act of 2004 mandated use of SATs and many of them are presented to analysts in intelligence training as methods for coping with their unavoidable "mindsets and biases" [8], [9], [10]. More recently, Intelligence Community Directive 203 on analytic standards, promulgated by the Office of the Director of National Intelligence (ODNI), states that analysts "must employ reasoning techniques and practical mechanisms that reveal and mitigate bias" (see Ref. [11], p. 2) by which they mean SATs. Variants of this approach have spread to many other nations [12], an excellent example of a phenomenon that sociologists

[1] This chapter is reprinted with permission from Mandel and Tetlock [1].

[2] Funding support for this work provided by the Canadian Safety and Security program projects [2] and Department of National Defence project 05da (Joint Intelligence Collection and Analytic Capability).

dub "institutional isomorphism." The SAT paradigm has spread not because there is evidence it works, but because influential professionals in the most powerful organization have endorsed it and no one wants to fall behind prevailing norms of best practices. In these environments, pressures for interoperability can easily trump systematic searches for optimal design, resulting in suboptimal cross-organizational learning.

1.3 CRITIQUE OF THE CURRENT APPROACH

The IC's current approach to judgment correctives is flawed for several reasons. We focus here on those that apply to the IC's general approach to judgment correction and do not descend into the weeds to critique individual SATs. Given space constraints, we condense our arguments into two areas of critique: core organizational limitations and core conceptual limitations. These areas are related, and have a common denominator in the IC's slow uptake from judgment and decision science, which followed from its commitment to an incidental approach, or lack of interest in pursuing a sustained, programmatic, and scientific approach to tradecraft innovation. We briefly address that common denominator before turning to the two areas of critique.

1.3.1 The Incidental Approach to IC Innovation

The IC's current approach to judgment correctives emerged from the attention of a handful of diligent analysts to specific problems they encountered in the practice of intelligence from the 1940s to 1980s. For instance, Kent's stubborn preoccupation with improving the fidelity of communications of uncertainty estimates was affected by his direct experience with a policymaker who was unsure of the meaning of the expression, "serious possibility," that appeared in a 1951 National Intelligence Estimate on the probability of a Soviet invasion of Yugoslavia that year [3]. When Kent asked his colleagues on the Board of National Estimates what they thought the term meant, he got answers ranging from 1:4 to 4:1 odds, which Kent described as jolting. Similarly, Heuer's interest in intelligence tradecraft – and "alternative analysis," in particular – was sparked by his involvement in the case of Soviet KGB defector Yuri Nosenko and his conclusion that the US IC made inadequate effort to consider alternative explanations for a string of suspicious events that seemed to support the conclusion that Nosenko was a KGB disinformation agent [13].

These tradecraft mavericks deserve credit for their trailblazing efforts to improve the practice of intelligence analysis. However, their examples also lay bare the adverse consequences of an *ad hoc*, character-driven approach to developing tradecraft. Critically, none of these tradecraft developers had advanced expertise in judgment and decision science. For example, although Heuer was well read in literature on higher order cognition, he did not pursue it at a professional or even post-graduate level, and he was not trained in research methods and statistical analysis. It is therefore unsurprising that he did not subject his methods – notably the Analysis of Competing Hypotheses (ACH) technique – to experimental tests of whether they actually improved judgment in measurable ways.

1.3.2 Organizational Limitations

Testing hypotheses is fundamental to both basic and applied sciences. Even our best ideas need to be put to rigorous empirical tests because most good ideas still fail. Mandel [14] recently argued that the IC's approach to tradecraft development follows what he called the goodness heuristic. Using this heuristic, if, upon mental inspection, an idea such as an imagined SAT for debiasing judgment seems good, then one should act on it as if it were in fact good because it probably is good. The goodness heuristic, which rests on a very likely excessively optimistic prior probability for ideational success, therefore takes Kahneman's [15] WYSIATI (What-You-See-Is-All-There-Is) principle to the next level by elucidating its implications for action by individuals and organizations.

Yet, as any seasoned scientist knows, not only do good ideas need to be rigorously tested, they need to be tested using multi-task and multi-benchmark methods [16]. There also should ideally be a diverse pool of ideas being tested by independent clusters of researchers, and among those clusters there must be a healthy sense of competition in epistemic tournaments, whether organized or *ad hoc* [17]. This is vital because scientists, as theorists, can become prisoners of their preconceptions all too easily [18]. Moreover, scientists, like all individuals, pursue goals other than purely epistemic ones [19]. It is vital, therefore, that scientists' ideas and key findings be subject to peer scrutiny.

Those who shaped the IC's current approach to judgment correctives varied in their commitment to testing ideas scientifically. Heuer, who had the greatest direct impact on the SAT approach to judgment correctives, questioned the value of science in adjudicating on the merits of proposed corrective methods. In an August 15, 2010 response to suggestions posted on an online discussion of the International Association for Intelligence Education that his ACH technique be empirically tested, Heuer wrote:

> *Can't we have confidence in making a commonsense judgment that going through the process of assessing the inconsistency of evidence will generally improve the quality of analysis? Similarly, can't we have confidence in making a commonsense judgment that starting the analysis with a set of hypotheses will, on average, lead to better analysis than starting by looking at the pros and cons for a single hypothesis? Do we really need an empirical analysis of these two points? Is it really feasible to do a high-quality empirical analysis of the effectiveness of these two points?*[20]

He also expressed reservations about the feasibility of experiments to test methods such as ACH, concluding, "If the empirical testing of my two claims about the value of ACH doesn't replicate exactly how ACH is (or should be) used in the Intel Community, I would be inclined to ignore it and stick with my commonsense judgment."

It is ironic that one of the IC's foremost tradecraft contributors, who stressed the importance of combatting confirmation bias, would take this stand. Yet the inconsistency should not shock us. The double standard – intuition is fine for me, but not for you – is simply more anecdotal evidence of the well-documented bias blind spot, the tendency to perceive biases in others' thinking and judgments more easily than in one's own [21].

We do not blame Heuer and others for exhibiting what most of us exhibit to varying degrees, but his stance highlights a consequence of the IC's decision over much of its history to invest very little in improving judgment quality through science, while investing heavily in collections technology. Over the last decade, the US IC has changed this approach and now funds the Intelligence Advanced Research Projects Activity (IARPA), which is programmatic, engaging large numbers of scientists from industry and academia, and which has led to important scientific advances that hold promise for improving intelligence products. Whether these advances can be effectively integrated into the analytic training and workflows of intelligence organizations remains to be seen.

1.3.3 Conceptual Limitations

The IC's traditional approach to analytic tradecraft has also fostered conceptual setbacks. While a heavy emphasis is placed on the mitigation of cognitive biases, virtually no attention is given to the problem of imprecision and unreliability caused by "noisy" unsystematic error [10]. Moreover, cognitive biases are conceptualized as unipolar phenomena needing to be reduced rather than as bipolar phenomena in which bias reduction strategies would require knowing where one was starting from, both in terms of direction and magnitude. Consequently, undue faith has been placed in assumptions regarding what types of biases needed to be corrected. For instance, whereas overconfidence is seen as problematic and attention is drawn to it in analytic training, the polar-opposite bias, underconfidence, is virtually ignored. However, recent studies

show evidence of underconfidence in strategic intelligence forecasts [22], [23] and in intelligence analysts' probability judgments in experimental tasks [24].

When we look at the research literature on how people cope with accountability demands [25], we worry that the IC's indiscriminate injunctions to beware of overconfidence will mainly yield indiscriminate response-threshold shifts – and the mirror-image bias of underconfidence. The net effect will be to further water down the informativeness of intelligence assessments for decision makers with excessive uncertainty. Similarly, the main effect of broad-brush warnings about confirmation bias might well be to induce endless second-guessing, to the point of analysis paralysis. Ultimately, the unipolar view of cognitive bias has allowed the IC to conveniently skirt value-laden, vexing questions about how bias reduction tradeoffs should be resolved.

The IC's error-neglect blind spot is equally troubling. Not only has the IC not taken proactive measures to minimize noise in intelligence judgments, noise neglect signals that the IC has not carefully considered how the very techniques they promote to minimize bias might amplify noise [10]. Yet the weakly defined multi-step processes that most SATs represent are no less than covert greenhouses for noise production. While giving the appearance of a standardized judgment-support process, SATs actually leave a long list of implementation decisions to analysts. How much agreement is there among analysts on such decisions? How reliably do the same analysts make these decisions over time? The few extant studies do not inspire optimism. For example, analysts asked to judge the probability of information accuracy on the basis of Admiralty code ratings of source reliability (i.e., A – F) and information credibility (i.e., 1 – 6) were unreliable when the two ratings were incongruent in ordinal value, and inter-analyst agreement plummeted as scale incongruence increased (see Ref. [26], Annex D).

In comparison to the Admiralty code, SATs like ACH create vast opportunities for inconsistency to flourish. To take just one example, consider the engine of ACH, which involves listing evidence in rows, hypotheses in columns, and then assessing the degree of consistency in each cell of the matrix. The meaning of consistency is left up to the analyst to interpret. One might treat it as the probability of the evidence given the hypothesis, while another might treat it as the inverse of that probability. Another still might assess whether the hypothesis necessarily follows from the evidence or vice versa, while yet another might run the test but with plausibility substituting for necessity. Perhaps the most common approach is to judge the representativeness of one to the other. In that case, and not without a touch of irony, ACH would be promoting the use of the representativeness heuristic under the guise of a debiasing strategy.

1.3.4 Correcting the IC's Current Corrective Approach

Both the organizational and conceptual limitations of the IC's approach to judgment correctives, in particular, and analytic tradecraft, in general, stem from its *ad hoc*, unscientific and character-driven nature. For the IC to develop effective correctives, it should abandon the complacent strategy of waiting for the next Kent or Heuer to spontaneously arise. The IC needs a diverse infusion of ideas from scientists outside the IC. It needs those scientists not only to put forward their best ideas, but also to test them in rigorous experiments or experimental tournaments. The IC should take the most promising results and work with scientific teams to transition the ideas into analytic processes. Those teams should also work with their IC counterparts to devise rigorous ways of trialing those processes, and the results of those trials should be taken seriously. What might work in an IARPA tournament, might not work so well in practice. If not, then reasons for variance in efficacy should be examined. Is the original idea doomed to transition failure, or was the transition strategy flawed but correctable?

The IC also should abandon the assumption that analytic judgments made in the absence of SATs must be intuitive and flawed. They should further banish the corollary view that although a SAT might not be perfect, it's better than nothing. The first assumption is certainly wrong and the second is probably wrong too. While intuitive processes enter into analysts' judgments, surely so can deliberative thought. SATs foster the illusion that intuition is driven from the judgment process. In reality, it is likely transferred to the process of

conducting the SAT exercise itself. The effects of such transfer can be far from banal. For instance, SATs might disrupt good deliberative reasoning about the substantive issues. They might bolster undeserved confidence in the accuracy and logical coherence of analysts' judgments. And they might foster IC complacency through the belief that corrective measures are sound and sufficient. For example, Mandel, Karvetski, and Dhami [27] report that intelligence analysts who were trained in ACH and who were instructed to use ACH to solve a probabilistic hypothesis-testing task were significantly more susceptible to coherence-violating unpacking effects [28] than a control sample of analysts from the same cohort who were not trained in ACH and who were left to their own reasoning devices.

Finally, the IC should broaden its horizons and start thinking beyond the analyst. All SATs share a focus on supporting the analyst, whether individually or in teams. Yet no attention has been given to how intelligence organizations might improve the accuracy of assessments through a range of post-analytic means such as recalibrating probabilistic judgments to correct for observable biases and aggregating judgments to boost signal-to-noise ratios through error cancellation and performance-sniffing methods. Recalibrating forecasts to make them more extreme has been shown to improve calibration in IARPA's "ACE" geopolitical forecasting tournament [29], [30] and in actual strategic intelligence forecasts [22]. Likewise, recalibration methods that "coherentize" probability judgments by forcing them to respect one or more axioms of probability calculus, such as additivity and unitarity, can improve accuracy [31]. The IC could also leverage decades of research on the benefits of statistically aggregating probability estimates. Taking an unweighted arithmetic average of multiple estimates is a highly effective method of error cancellation [32]. More sophisticated aggregation methods that exploit individual differences in coherence [31], [33], [34] or other measurable aspects of performance [35] also hold promise for the IC. Indeed, Mandel et al. [27] found that analysts' judgment accuracy was substantially improved by first "coherentizing" and then aggregating their judgments.

To accelerate the discovery process, the IC should also take steps to systematically monitor the accuracy of its products. This will reveal the types of corrective actions most needed, and it can also shed light on factors that predict judgment accuracy. The results may be counterintuitive and impossible to predict from theory. For instance, contrary to intuitive expectation, topic-related expertise among cancer research experts did not predict better accuracy in forecasting the reproducibility of cancer trial results, but expertise defined in terms of publication impact (h-index) did [36]. Likewise, Tetlock [37] found that political experts working inside their self-described domain of competence were no more accurate than experts working outside their domain in a geopolitical forecasting tournament. Ferreting out the factors that could be used in performance-sniffing weighting methods will take time and research effort, but these and other post-analytic interventions could significantly boost the IC's judgment accuracy in years to come. The IC only needs to reduce the probability of a trillion-dollar mistake by a tiny amount to justify multi-million-dollar research investments.

1.4 REFERENCES

[1] Mandel, D.R., and Tetlock, P.E. (2018). Correcting judgment correctives in national security intelligence. *Frontiers in Psychology* 9:2640. https://doi.org/10.3389/fpsyg.2018.02640.

[2] Canadian Safety and Security Program Projects CSSP-2016-TI-2224 (*Improving Intelligence Assessment Processes with Decision Science)* and CSSP-2018-TI-2394 (*Decision Science for Superior Intelligence Production)*.

[3] Kent, S. (1964). Words of estimative probability. *Studies in Intelligence* 8 (4):49-65.

[4] Scoblic, J.P. (2018). Beacon and warning: Sherman Kent, scientific hubris, and the CIA's Office of National Estimates. *Texas National Security Review* 1 (4): 99-117.

[5] Heuer, R.J., Jr. (1999). *The Psychology of Intelligence Analysis*. Washington DC: Central Intelligence Agency, Center for the Study of Intelligence.

[6] Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.

[7] Heuer, R.J., Jr., and Pherson, R.H. (2008). *Structured Analytic Techniques for Intelligence Analysis.* Washington DC: CQ Press. Developed by Pherson Associates, LLC.

[8] Marchio, J. (2014). Analytic tradecraft and the intelligence community: Enduring value, intermittent emphasis. *Intelligence and National Security*, 29 (2):159-183.

[9] Coulthart, S.J. (2017). An evidence-based evaluation of 12 core structured analytic techniques. *International Journal of Intelligence and CounterIntelligence,* 30 (2):368-391.

[10] Chang, W., Berdini, E., Mandel, D.R., and Tetlock, P.E. (2018). Restructuring structured analytic techniques in intelligence. *Intelligence and National Security* 33 (3):337-356.

[11] Office of the Director of National Intelligence. (2015). *Intelligence Community Directive 203: Analytic Standards.* Washington DC: Office of the Director of National Intelligence. Retrieved from https://fas.org/irp/dni/icd/icd-203.pdf

[12] Butler, F.E.R., Chilcot, J., Inge, P.A., Mates, M., and Taylor, A. (2004). *Review of Intelligence on Weapons of Mass Destruction.* The Butler Review, HC 898. London, UK. Retrieved from http://news.bbc.co.uk/nol/shared/bsp/hi/pdfs/14_07_04_butler.pdf.

[13] Heuer, R.J., Jr. (1987). Nosenko: Five paths to judgment. *Studies in Intelligence* 31 (3):71-101.

[14] Mandel, D.R. (2019). Can decision science improve intelligence analysis? In: *Researching National Security Intelligence: Intelligence: Multidisciplinary Approaches*, Coulthart, S., Landon-Murray, M., and Van Puyvelde, D. (Eds.), 117-140. Washington DC: Georgetown University Press.

[15] Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

[16] Mellers, B.A., Baker, J.D., Chen, E., Mandel, D.R., and Tetlock, P.E. (2017). How generalizable is good judgment? A multi-task, multi-benchmark study. *Judgment and Decision Making,* 12(4):369-381.

[17] Tetlock, P.E., Mellers, B.A., and Scoblic, J.P. (2017). Bringing probability judgments into policy debates via forecasting tournaments. *Science* 355(6324):481-483.

[18] Tetlock, P.E., and Henik, E. (2005). Theory- versus imagination-driven thinking about historical counterfactuals: Are we prisoners of our preconceptions? In: *The Psychology of Counterfactual Thinking*, Mandel, D.R., Hilton, D.J., and Catellani, P. (Eds.), 201-244. New York, NY: Routledge.

[19] Mandel, D.R., and Tetlock, P.E. (2016). Debunking the myth of value-neutral virginity: Toward truth in scientific advertising. *Frontiers in Psychology,* 7:451.

[20] Heuer, R.J., Jr., August 15, 2010 email correspondence sent to the International Association for Intelligence Education.

[21] Pronin, E., Lin, D.Y., and Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin,* 28 (3):369-381.

[22] Mandel, D.R., and Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences,* 111(30):10984-10989.

[23] Mandel, D.R., and Barnes, A. (2018). Geopolitical forecasting skill in strategic intelligence. *Journal of Behavioral Decision Making,* 31(1):127-137.

[24] Mandel, D.R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Frontiers in Psychology*, 6:387.

[25] Lerner, J.S., and Tetlock, P.E. (1999). Accounting for the effects of accountability. *Psychological Bulletin,* 125(2):255-275.

[26] Mandel, D.R. (2018). *Proceedings of SAS-114 Workshop on Communicating Uncertainty, Assessing Information Quality and Risk, and Using Structured Techniques in Intelligence Analysis*. NATO Meeting Proceedings. Brussels, Belgium: NATO STO.

[27] Mandel, D.R., Karvetski, C., and Dhami, M.K. (2018). Boosting intelligence analysts' judgment accuracy: What works, what fails? *Judgment and Decision Making,* 13(6):607-621.

[28] Tversky, A., and Koehler, D.J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review,* 101(4):547-567.

[29] Baron J., Mellers, B.A., Tetlock, P.E., Stone, E., and Ungar, L.H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis,* 11 (2):133-145.

[30] Turner, B.M., Steyvers, M., Merkle, E.C., Budescu, D.V., and Wallsten, T.S. (2014). Forecast aggregation via recalibration. *Machine Learning*, 95(3):261-289.

[31] Karvetski, C.W., Olson, K.C., Mandel, D.R., and Twardy, C.R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis,* 10(4):305-326.

[32] Clemen, R.T., and Winkler, R.L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis,* 19(2):187-203.

[33] Predd, J.B., Osherson, D.N., Kulkarni, S.R., and Poor, H.V. (2008). Aggregating probabilistic forecasts from incoherent and abstaining experts. *Decision Analysis,* 5(4):177-189.

[34] Wang, G., Kulkarni, S.R., Poor, H.V., and Osherson, D.N. (2011). Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis,* 8 (2):128-144.

[35] Cooke, R.M., and Goossens, L.L.H.J. (2008). TU Delft expert judgment database. *Reliability Engineering System Safety*, 93(5):657-674.

[36] Benjamin, D., Mandel, D.R., and Kimmelman J. (2017). Can cancer researchers accurately judge whether preclinical reports will reproduce? *PLoS Biology* 15 (6):e2002212. https://doi.org/10.1371/journal.pbio.2002212.

[37] Tetlock, P.E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.



Chapter 2 – MITIGATING RISK IN THE ANALYTIC WORKFLOW: A UK PERSPECTIVE

DI Futures and Analytical Methods
Ministry of Defence
UNITED KINGDOM

2.1 INTRODUCTION

Across Defence in the UK, decisions are taken for many different reasons and at different levels, from military commanders in operational theatres to cross-government strategic policy. What they have in common in almost all cases, however, is a reliance on information that may be incomplete, conflicting, unreliable and/or unknowable. Further uncertainties are introduced by the processing, interpretation, and exploitation of such information by humans and computers. The role of Defence Intelligence (DI) in the UK is to reduce uncertainty for decision makers so that the optimal courses of action are more likely to be taken, or policies adopted.

This chapter presents a framework for the evaluation of uncertainties throughout the analytic process in order to maximise value to decision makers and reduce the risk of intelligence failures. It categorises different types of uncertainty and outlines mitigation measures, enabling identification of areas where no interventions exist. Connections with SAS-114 research and contributions are highlighted to demonstrate where such work has influenced DI's approach to mitigating analytic risk.

2.2 DEFENCE INTELLIGENCE IN THE UK

Defence Intelligence is the UK Government's primary producer of strategic intelligence concerning the defence of the UK and its overseas interests. Alongside joint and single service military intelligence organisations, DI also contributes to the production of operational and tactical intelligence for military commanders. While historically associated with all-source intelligence assessment, DI now comprises staff across a broad range of collection, analysis, assessment, and enabling activities. These include civilian and military specialists in all intelligence disciplines, working in several geographic locations.

2.2.1 Threat Assessment and Risk Assessment in DI

Defence Intelligence fulfils its role of reducing uncertainty for decision makers by assessing threats posed by foreign adversaries across a range of regional and thematic topics. Defining threat as 'capability + intent + activity', DI is responsible for determining the nature of current and future threats and making judgements about the probability of different threats being realised. It contributes to government-level prioritisation frameworks to shape DI's priorities, but does not routinely evaluate costs to UK interests that may result if a threat is realised. It is specifically prohibited from recommending policy or mitigation against threats.

If risk is defined as the probability of a threat multiplied by the associated cost or impact, DI does not undertake risk assessment with respect to foreign adversaries' actions. This is the responsibility of the decision makers who evaluate the assessed probability against the assessed costs or impacts, and decide what courses of action are required, if any, in order to reduce or eliminate the risk to UK interests. However, DI should reduce decision makers' uncertainty by providing robust, justifiable probabilistic threat assessments on which the decision makers' risk assessments can be based.

2.3 THE ANALYTIC RISK FRAMEWORK

2.3.1 Intelligence Failure as a Risk

Intelligence failures occur when shortcomings in the direction, conduct, communication, or use of intelligence analysis and its products result in physical, reputational, financial and/or political damage to intelligence producers, intelligence users and/or third parties. These shortcomings may be manifested as incorrect analytic conclusions, and intelligence failures are often portrayed as analysts 'getting it wrong'. However, it is important to consider that missed opportunities for analysis, misunderstandings, and poor use of intelligence in decision making can also be types of intelligence failure as these can also result in inappropriate policies and/or actions. The analytic risk framework places intelligence failure as the risk that needs assessing to determine if mitigation is necessary.

2.3.2 Assessing the Risk of Intelligence Failure

Intelligence failures, as defined above, are inevitable due to the complexities of dealing with uncertainties throughout the intelligence cycle, but their frequency and impact can be reduced if the risk is understood. The analytic risk framework separates intelligence failures into five categories, each related to distinct sources of uncertainty during the analytic process: requirement, quantity of information, quality of information, judgement, and output. Threats contributing to the risk of intelligence failure have been described within subcategories. By systematically evaluating the probability of threats being realised and the resulting costs, organisations can ascertain whether they have sufficient mitigation against intelligence failures.

While it is clearly preferable to reduce the probability of threats occurring in the first place, the inevitability of intelligence failure means that mitigation aimed at reducing costs of a failure may be just as appropriate, especially when considering factors outside the intelligence producer's control.

The consequences, or costs, associated with intelligence failures range from temporary or localised impacts to longer-term ramifications affecting significant numbers of people. The extent to which these costs materialise will vary according to the specific nature, scale, and severity of the intelligence failure.

2.3.2.1 Cost to the Analyst

The analyst responsible for the intelligence may experience reputational damage and a loss of trust from peers and seniors, as well as frustration due to the time wasted. In more extreme cases, the individual may be subject to legal proceedings or longer-term career damage. They could also suffer from guilt, anxiety, or more severe mental health issues if there were further consequences impacting other parties.

2.3.2.2 Cost to DI

The costs to DI associated with intelligence failures include reputational damage, potentially reducing its value to decision makers. Resources would have been wasted in the production of the intelligence with knock-on effects elsewhere in the organisation. It could also face legal proceedings or financial penalties depending on the nature and severity of the failure.

2.3.2.3 Cost to the UK Government

At a UK Government level, potential costs associated with intelligence failure also include wasted resources, legal proceedings and financial penalties. However, the impacts here extend to domestic political damage, international reputational damage, physical damage to infrastructure and assets, and injury or death for UK citizens or allies if intelligence is incorrect, misunderstood, or misused.

2.3.2.4 Cost to Wider Society

Wider society may lose trust in security and intelligence services, law enforcement, or politicians, potentially damaging national institutions. Members of the public may be exposed to injury, death, or destruction of property. The financial implications for taxpayers could also be significant in terms of legal proceedings and reparation for any resulting physical or human costs, as well as wasted government resources producing the intelligence.

2.3.3 Application to DI

This chapter evaluates DI's position against this framework in terms of current mitigation practices and how they link with SAS-114 research. Each section contains a table outlining the application of the framework to one of the five categories, including the threats in each subcategory, with explanatory text for context. The subsequent narrative expands on how DI mitigates against threats, including whether this is focussed on reducing the probability of a threat occurring or the cost if it does. It is important to note here that the probability and costs are conceptual amounts only; no attempt has been made to quantify either. This enables consideration of different types and magnitudes of costs for all threats. Each section concludes with a discussion of how SAS-114 research has contributed to DI's understanding or practices, or may do so in the future.

Other organisations can use the framework to evaluate their own position with regard to mitigation practices. This may require some modification to account for different remits within organisations outside DI. For example, some intelligence organisations require analysts to include recommendations for action along with their analysis, which may necessitate the inclusion of further threats at the output stage. However, most threats listed will be applicable to any intelligence production organisation.

2.4 REQUIREMENT

2.4.1 Outline

Table 2-1: Assessment of Analytic Risk Associated with Requirement.

Requirement				
Subcategory	**Threats**	**Mitigation Focus**	**Mitigation**	**SAS-114**
Purpose	Misunderstood rationale for requirement; Relevant customers not (all) identified; Arbitrary or misunderstood deadlines.	Probability and Cost	• Customer liaison • Prioritisation frameworks • Dedicated requirements managers • Customer education on capabilities • Question refinement • Deadlines	[1], [2], [3], [4], [5], [6], [7], [8]
Nature	Requirement not achievable; Requirement is unclear or ambiguous; Requirement has been misinterpreted.	Probability and Cost		
Prioritisation	Priorities have been incorrectly set due to misunderstanding or deception; Resources not available to meet requirement; Resources allocated to lower priority tasks.	Probability and Cost		

Intelligence failures related to requirement arise from uncertainties over what is being asked, by whom and why, and the extent to which this is truly understood by both customers and producers. Customers often do not fully understand DI capabilities and may ask questions that are based on assumptions or misunderstandings, or are unachievable in the requested time. Similarly, tasking authorities within DI may misinterpret questions or not recognise the underlying purpose that the request is addressing. If the requirement is not truly understood, resources may be allocated inappropriately, including numbers of staff, expertise, and time allowed, resulting in intelligence failures from the wrong issues being addressed, or questions being answered too late or not at all. This would be exacerbated if priorities had been misidentified by accident or subterfuge. In addition, some customers may have similar requirements, which, if not identified, could lead to wasted resources through duplicated effort or missed opportunities to influence decisions.

2.4.2 Mitigation

Defence Intelligence's mitigation against these types of intelligence failure largely fall within the responsibility of staff in tasking, prioritisation, outreach, and customer liaison roles, although all analysts have a duty to clarify the requirement they are addressing as far as they are able. Mitigation includes working with prioritisation frameworks shared across UK Government and dedicated requirements managers who work with different customers to ensure requirements are captured accurately, with realistic and justifiable deadlines. Opportunities to educate customers are present in the form of roadshows, familiarisation visits, embedded liaison officers, and less formal desk level events. An established three-step question refinement method is taught to analysts in early career training and reinforced through guidance material and tradecraft advice.

These mitigation practices are predominantly aimed at reducing the probability of a threat being realised, by ensuring that customers and producers have a shared and accurate understanding of what is required and what is achievable, and that this is suitably factored into decisions about priorities. The prioritisation frameworks and requirements managers also provide a mechanism for reducing the cost of any failure by minimising the potential for unnecessary or duplicated work and providing robust justification for resource allocation to minimise reputational damage.

2.4.3 SAS-114 Contribution

Research and discussions at the SAS-114 technical team meetings have predominantly informed DI's thinking with respect to how intelligence requirements are framed and how this influences how they are interpreted. In particular, the importance of clarifying the desired outcome rather than just accepting a question was highlighted in presentations about information utility and the risk game. In both cases, the research showed that the perceived or actual desired outcome affected what information was used and the approach taken. Similarly, other presentations touched on how the type of task needs to be considered in order to determine the most appropriate method. These indicated that DI should ensure that sufficient emphasis is placed on understanding customer goals rather than stated requirements, as it may be that resources are wasted trying to find a single definitive answer when a more general reduction of likely options is sufficient.

Research examining the way intelligence requirements are presented has also been of interest. Dynamic questioning and prediction markets would both significantly change the way requirements are formed and managed. They have potential benefits with regards to improving the understanding of a requirement and the ability to address it, but both have some technical, organisational, and cultural obstacles that limit their adoption within DI.

2.5 QUANTITY OF INFORMATION

2.5.1 Outline

Table 2-2: Assessment of Analytic Risk Associated with Quantity of Information.

Quantity				
Subcategory	**Threats**	**Mitigation Focus**	**Mitigation**	**SAS-114**
Existence of information	Information does not exist to meet requirement in time; Alternative and novel sources overlooked.	Probability	• Repositories • Collaborative platforms • Mandated use of metadata • Search tools • Data management practices • Data management specialists • Basic training (sources of information) • Basic training (systems) • Collection plans • Collection management teams	[2], [3], [9], [10]
Collection	Information cannot be collected in time to meet requirement; Collection opportunities missed; Collection bias.	Probability		
Availability of information	Information exists but not accessible or usable due to system functionality, classification and/or data management practices; Information exists but not known about or exploited.	Probability		
Sifting and prioritisation	Inability to find relevant information when high data volume.	Probability		

Intelligence failures originating from the quantity of information available relate to uncertainties about whether analysts have, or can collect, sufficient information to answer an intelligence question and if they can access it within the time allocated to the task. Almost all intelligence analysis is based on incomplete information, which introduces uncertainty by increasing the reliance on assumptions rather than evidence. Assessments that minimise the use of assumptions better meet DI's role of reducing uncertainty for decision makers and are therefore more likely to assist in making effective decisions. In most instances, the use of assumptions does not preclude an assessment being formed, especially if they are not critical assumptions that would undermine the analytic conclusions if proved false. However, there are extreme cases in which a lack of information reaches an intolerable level, making any assessment little more than an unsubstantiated guess based on the analyst's individual experience. These can include hypotheses about past events, where collection opportunities have already been missed, or extremely volatile future scenarios that are difficult to predict. Collection opportunities can be exploited to increase the amount of information, including using new or alternative sources, but these are subject to the same uncertainties as existing information in terms of timeliness and reliable access.

It is not just a lack of information, however, that can cause uncertainty. The rapidly increasing amount of information available to analysts through the internet or other digital sources means that finding the best information amongst the noise can be challenging. Analysts also need to guard against collection bias, where

one hypothesis appears to be more supported than another solely because of the relative volume of information. The ability to sift and prioritise is critical to avoiding intelligence failures arising from significant information being missed or making assessments based on inappropriate information.

2.5.2 Mitigation

Defence Intelligence's means of mitigation against these types of intelligence failure are predominantly technological in nature. They include the provision of information repositories and collaborative platforms that enable analysts to search for and share information with appropriately cleared colleagues. Data management practices, such as the use of metadata, are employed to ensure access to information, and some specialist areas have dedicated data management teams. However, there remains a risk where data is held on different IT systems. Training is provided on the different systems, but much is still reliant on experience, supervision, and individual effort to maximise the discovery of and access to relevant information.

In addition to mitigation focussed on finding and accessing existing information, DI has teams responsible for collection management. These teams provide expert advice on different collection capabilities and can work with analysts to build collection plans against intelligence topics, including less well-known capabilities.

All of these mitigation practices are aimed at reducing the probability of a threat being realised, by enabling analysts to gather and exploit as much relevant information as possible to answer their task.

2.5.3 SAS-114 Contribution

Since most information collected is in a digital format, the responses are largely technological in nature and, therefore, on the periphery of the remit of the SAS-114 panel. Dynamic questioning offered an interesting conceptual method for more targeted searching of information (related to the identification of the desired outcome described earlier), but significant technical barriers limit implementation of the demonstrated software in DI. Similarly, presentations on measures of information and information utility provided useful ideas for how to identify and prioritise useful evidence, although their primary value related to information quality (below).

2.6 QUALITY OF INFORMATION

2.6.1 Outline

Table 2-3: Assessment of Analytic Risk Associated with Quality of Information.

Quality				
Subcategory	**Threats**	**Mitigation Focus**	**Mitigation**	**SAS-114**
Individual source reliability	Source is untrustworthy/ unreliable; Human sources with uncertain or questionable credentials and motivations; Technical sources with inappropriate parameters and/or prone to errors.	Probability	• Source grading (some sources) • Basic training (sources of information) • Structured analytic techniques	[1], [2], [3], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21]
Individual source credibility/ content	Information is incorrect, inconsistent or unsubstantiated; Potential for denial and deception.	Probability		

Quality				
Subcategory	**Threats**	**Mitigation Focus**	**Mitigation**	**SAS-114**
Information relevance	Information used is not pertinent to requirement; Information is dismissed erroneously; Information relevance not reviewed over time.	Probability		
Information significance	Impact of information is under or over estimated.	Probability		
Aggregation	Analyst unconsciously favours certain types of information; Inconsistent handling of multiple uncertainties; Inconsistent treatment of agreement or disagreement within information.	Probability		

Intelligence failures stemming from the quality of information arise from incorrect value being placed on information in answering a specific intelligence question. The extent to which information is reliable, relevant, and significant are all subjective judgements for the analyst, as is the reliability of the source itself. Consideration should be given to the reliability of both the source and the information, as a trustworthy source could still contain accidentally or deliberately false information. This subjectivity causes uncertainty in terms of how much weight should be assigned to different pieces of information, especially when there is conflicting information, as this can vary significantly between analysts. The same piece of information could have different values for different questions, and at different times, so it is important that it is evaluated in the context of a specific task. Uncertainty may be further increased by how analysts aggregate multiple pieces of information with differing measures of reliability and credibility as this may be inconsistent and/or subject to cognitive biases. Inappropriate and/or inconsistent judgements as to the value of information can undermine the accuracy of assessments and therefore their value to customers.

2.6.2 Mitigation

Some information used by DI analysts incorporates a source grading. In these cases, the producer of that information has determined how reliable it is based on their knowledge of the source. This removes that responsibility from the analyst, ostensibly increasing consistency. Analysts receive basic training on the strengths and weaknesses of different sources, enabling them to consider how well that source can contribute to their understanding of an intelligence issue. Introductory analytic training also covers structured analytic techniques that can be used to evaluate information systematically to minimise inconsistencies and mitigate bias.

All of these mitigation practices are aimed at reducing the probability of a ‘threat’ being realised by enabling analysts to make more reasoned and consistent judgements on the value of information.

2.6.3 SAS-114 Contribution

SAS-114 research and discussions related to quality of information can be divided broadly into three themes. Firstly, in terms of characterising uncertainty in this area, collation of different defence and security standards for uncertainty highlighted large variance in what is meant by different elements of uncertainty but

was able to draw out common characteristics in order to clarify thinking. Other research highlighted the need to account for lexical context when undertaking automated text extraction in order to accurately capture information content.

Secondly, multiple presentations addressed the effects of source reliability and information credibility on analysts' application and evaluation of information. These highlighted variance and inconsistency in how source gradings were used and interpreted, even when done so in a structured framework such as an accepted NATO standard or a structured analytic technique. The results of one set of experiments indicated that the type of information was more important than an assigned quality value [16], [20]. This suggests that analysts tend to make independent valuations of information that resonate with what they are familiar with, possibly supporting DI's emphasis on source education over explicit gradings.

Thirdly, information relevance and significance were discussed in numerous presentations related to information usefulness and information gain. These provided a useful theoretical basis for considering how to improve analysts' ability to identify the most valuable information for answering their questions, both in terms of existing information and intelligence gaps to be collected against.

Overall, research in this area was thought provoking for looking at existing practices within DI and whether they can be improved upon. Certain findings would impact standards and practices over which DI has no control or authority, but there is potential for lower-level adjustments that could improve how analysts think about and deal with information quality.

2.7 JUDGEMENT

2.7.1 Outline

Table 2-4: Assessment of Analytic Risk Associated with Judgement.

Judgement				
Subcategory	**Threats**	**Mitigation Focus**	**Mitigation**	**SAS-114**
Approach	Inappropriate choice of method or approach; Failure to consider different approaches; Inconsistent application of terms and tools over time and between colleagues; Inability to replicate in the future (accounting for changes over time).	Probability	• Structured analytic techniques (training, guidance material) • Analytic consultancy service • Taxonomy for selecting structured analytic techniques • Training (analytic concepts) • Analytic standards • Probability yardstick • Recruitment policies (limited)	[1], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [18], [19], [20], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36]
Decision making	Lack of evidence supporting conclusions; Inability to justify (rejected) conclusions; Illogical argumentation; Failure to consider different explanations; Insufficient objectivity.	Probability		
Information gaps	Assumptions not identified and evaluated; Information gaps and ambiguities not addressed.	Probability		

Judgement				
Subcategory	**Threats**	**Mitigation Focus**	**Mitigation**	**SAS-114**
Self-reflection	Failure to recognise and address bias in analysis or analyst; Over or underestimation of own abilities and/or experience; Analytic conclusions vary according to perceived consequences (fear of being 'wrong').	Probability and Cost	• Training (subject-specific) • Advocacy of intellectual audit trails • Promotion of challenge culture • Reviews of previous work (limited)	
Subject matter knowledge/ expertise	Insufficient subject matter knowledge; Overreliance on subject matter knowledge; Failure to challenge SME.	Probability		

Intelligence failures related to judgement can result from a wide range of uncertainties regarding how analysts think about and process information in order to make assessments. Thinking and reasoning are, by their very nature, subjective processes that will vary between analysts and over time. They are also usually internal processes, making them difficult to articulate or replicate. This variation and opacity generate uncertainty in terms of being able to justify and demonstrate how conclusions and assessments were reached. Unsound reasoning, insufficient consideration of alternative approaches and/or explanations, failure to use appropriate evidence, and hidden or poorly addressed assumptions increase the likelihood of incorrect conclusions.

Further uncertainty can be introduced by influences on those thought processes, both internal and external. An analyst's personal biases and view of their own abilities, knowledge, and experience may unduly influence their assessments, as may the perception of the consequences of that assessment, especially if it could have negative impacts. Similarly, too much emphasis may be placed on colleagues' knowledge if they are identified as 'experts'. While self-confidence and subject matter expertise can be positive influences on analysis, intelligence failures can occur when they go unchallenged, allowing unsubstantiated assumptions to persist with knock-on effects on accuracy and value of the resulting assessments.

2.7.2 Mitigation

The majority of DI's practices to mitigate intelligence failures related to judgement are focussed on equipping analysts with knowledge, skills, and tools to enable them to recognise and minimise the causes. This includes methods to increase objectivity, standardised approaches, and the externalisation of thought processes. All analysts receive introductory training on analytic concepts such as biases, fallacies and assumptions, and on a selection of structured analytic techniques. The latter are intended to assist in exposing biases, assumptions, and thought processes, enabling potential flaws to be recognised, challenged, and addressed prior to report publication by the analyst themselves, peers, or supervisors. For some analysts, they provide a more rigorous approach that improves the quality of the analysis; for others, they are purely for capturing thought processes for evidence and audit purposes. Analysts are also provided with a taxonomy to help them select the most appropriate structured analytic technique(s). In addition to the training, analysts also have access to an analytic consultancy service that provides bespoke advice, guidance material, and practical assistance to help individuals and teams structure their response to specific intelligence questions, often using a wider range of techniques than covered on the training. A tool for evaluating analytic confidence is in development, enabling analysts and supervisors to articulate different factors affecting how confident they are in a probabilistic judgement prior to report publication.

In early 2018, DI introduced a set of formal analytic standards that detailed what was expected of analysts when producing all-source intelligence reports. In addition to setting standards for analytic rigour, these include the requirement for analysts to be independent and objective, emphasising the importance of impartial use of evidence above self-interest or external pressure. The mechanisms for monitoring adherence to these standards are being established, so the impact cannot yet be determined.

Recruitment into analytic posts sometimes requires candidates to demonstrate functional analytic competence, using examples of previous experience. This may include an element of subject matter expertise regarding the intelligence topic they are likely to be working on. However, demonstration of functional competence or SME is not mandated, leaving some analysts more reliant on training and supervision, rather than innate ability or experience. These needs can sometimes be insufficiently met due to capacity and capability limitations, with the potential for a resulting negative impact on quality of judgement. Some subject-specific training is available but subject matter knowledge usually develops through experience after an initial immersion with background reading.

Most of these mitigation practices are aimed at reducing the probability of a threat being realised by improving analysts' ability to think in a more structured and systematic way about how they are reaching an assessment. However, some attempts are made to reassure analysts that they will be protected from blame and reputational or career damage if their assessment is later proven wrong but they can adequately demonstrate and justify how they reached it.

2.7.3 SAS-114 Contribution

This area is where the majority of SAS-114 research informs and influences DI's thinking. Numerous presentations added to the evidence base for understanding and improving how analysis is conducted, including testing the performance of some structured analytic techniques and performing contrarian analysis to challenge accepted thinking. Research was presented into different thinking styles and the impact this has when conducting different types of task. A significant number of studies looked specifically at how probabilities were managed and measured, from pure mathematical approaches of deductive logics and frequency formats to more subjective relative judgements. The nature of intelligence analysis in DI makes the latter more directly and immediately applicable, but mathematical theories enable more rigour when considering further mitigation against intelligence failures caused by insufficient understanding of probabilities. Initial results of an exercise measuring predictive accuracy in a stakeholder organisation highlighted the importance of such information in capturing evidence of what practices work and what need further interventions [22], [25].

Other presentations contributed to DI's understanding of more conceptual components of decision making, including entropy and information gain, quasi-rationality, reasoning profiles, and uncertainty handling frameworks. Common challenges were identified in stakeholder organisations regarding legacy approaches to analysis and how they impacted implementation of evidence-based processes, with discussion about examples of successful organisational changes. This included the introduction of new methods for generating intelligence assessments in the form of prediction markets, which overcome some of the issues leading to errors in judgement but would require DI to overhaul its approach to managing requirements and analysts.

Given the panel's remit, it is perhaps unsurprising that this is where research has been most productive. Many issues and topics were raised that would be of interest to DI if future opportunities exist to explore them.

2.8 OUTPUT

2.8.1 Outline

Intelligence failures connected to output relate to uncertainties over how the results of intelligence analysis are communicated (both formally and informally) and valued. A significant source of uncertainty related to output is how the content is interpreted, as ambiguity, vagueness, or misrepresentation can result in intelligence failures through accidental or deliberate misuse by decision makers. This will be exacerbated if the output is subject to intermediate analysis, such as summaries of reports being provided for senior decision makers or briefings delivered by someone other than the product author. The opportunity to influence decisions will be missed if products provide limited value to customers due to not meeting requirements, being too late, or not being accessible (classification or system limitations). Similarly, for different customers the inclusion or exclusion of supporting detail such as methodology or references can impact the value of a product, and the presentation of output can also cause intelligence failures if unappealing formats or poor quality control causes the content of products to be overlooked. In rare cases, intelligence may be dismissed if it does not contain the conclusions a decision maker expected or wanted.

The criteria used to measure output value internally can directly or indirectly influence the value to customers by prioritising potentially conflicting aims. For example, analysts may be rewarded for disseminating a large number of reports on time, rather than appreciating the impact that those reports should have, or an emphasis may be placed on accuracy that leads to analysts favouring vagueness over precision in assessments to avoid being proved wrong.

Table 2-5: Assessment of Analytic Risk Associated with Output.

Output				
Subcategory	**Threats**	**Mitigation focus**	**Mitigation**	**SAS-114**
Value (External/ customer)	Output is of limited or no value to customer as it fails to meet requirement and/or deadline; Output overlooked as it does not meet expectations.	Probability and Cost	• Customer engagement and education • Repositories • Collaborative platforms • Data management practices • Mandated use of metadata • Centralised dissemination team • Analytic standards • Writing standards • Product standards and templates • Production checklists	[4], [5], [7], [8], [11], [18], [22], [23], [25], [28], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46]
Accessibility	Customer cannot access output when required; Output is not discoverable to potential customers.	Probability		
Value (Internal)	Internal value systems undermine value to customers.	Probability		
Interpretation of conclusions	Insufficient distinction between fact/evidence, assumptions and assessment; Ambiguity or lack of clarity in language (including probabilistic judgements); Conclusions misinterpreted or distorted by intermediate analysis.	Probability and Cost		

Output				
Subcategory	**Threats**	**Mitigation focus**	**Mitigation**	**SAS-114**
Accuracy of conclusions	External and/or unforeseeable influences/events; Reason(s) for (in)accurate conclusions are not known or addressed.	Probability and Cost	• Pre-publication reviews • Probability yardstick • References • Audit trails • Reviews of previous work (limited)	
Presentation	Poor quality control; Inappropriate and/or unappealing formats; Unsuitable level of detail and/or exposure of methodology.	Probability		

2.8.2 Mitigation

Defence Intelligence's mitigation strategies against intelligence failures related to output are focussed on the areas over which they have most control. As with mitigation against failures associated with requirement, there is an element of customer engagement to ensure that customers know what information is available to them and how to access it, as well as regular requests for feedback from key customers. There have been attempts to measure the effect of intelligence products over the years, but none have been sustained. Information management practices, including the use of repositories, metadata, collaborative platforms, distribution lists, and centralised dissemination are also used to make output available to as broad an audience as possible.

A suite of standards covering analysis, writing styles and product formats (templates, etc.) is in place to describe expectations and requirements for products. In terms of output, the analytic standards require analysts to be cognisant of the clarity of their language, along with the relevance, timeliness, and auditability of products. This includes mandating the use of the PHIA[1] Probability Yardstick, a UK Intelligence Community standard that details approved language for communicating uncertainty to minimise the possibility of a customer interpreting probabilistic judgements differently to how the analyst intended. Formal procedures, such as checklists and peer review, and/or dedicated teams provide quality control prior to publication of products, with some areas of DI requiring specific qualifications for those with authority to release reports. Fully referenced audit trails demonstrating how conclusions were reached are mandated in some areas in order to provide a formal record of activities. Post-publication reviews of selected products are sometimes undertaken to identify good and bad practice, but these are not routine for most areas.

All of the mitigation practices are aimed at reducing the probability of a threat being realised by ensuring that output is as clear as possible about the results of the intelligence analysis, and that it reaches customers when needed. However, as potential intelligence failures in this area are more likely to occur due to actions outside DI's control, these practices also help reduce the cost by minimising DI's culpability if output is overlooked or misused.

2.8.3 SAS-114 Contribution

The majority of research and discussions within SAS-114 were concerned with output related to communication of uncertainty, particularly probabilistic language. Numerous challenges were highlighted,

[1] Professional Head of Intelligence Assessment – the UK Government's co-ordinating body for standards for the Intelligence Assessment profession.

including multiple, different sets of communication standards, inconsistent awareness of when and how to apply standards and the complicating factor of having to understand and account for varied customer preferences, requirements and tolerance for uncertainty. The importance of considering customer requirements in terms of style and format was also discussed in some presentations, especially given developments in digital presentation formats and on demand delivery.

In addition to communication, early results of research studying how analytic standards were perceived by analysts was presented with possible implications for the standards recently introduced in DI. Other stakeholders discussed how adherence to standards is monitored in their organisations, further informing the development of DI's quality assurance programme.

Research regarding communication of intelligence analysis is perhaps the most difficult to implement in DI as it is most heavily determined by external organisations, inside and outside the Ministry of Defence. As seen in other stakeholder organisations and discussed at length in SAS-114 meetings, there is no 'perfect' solution to communicating probability. DI must adhere to standards set by the UK Intelligence Community co-ordinating body, PHIA, and find ways to ensure analysts and customers understand these alongside those of allied counterparts.

2.9 SUMMARY

The analytic risk framework is a method of exploring how well an organisation is equipped to minimise the risk of intelligence failure in the intelligence analysis process. Categorising different types of risk enables an organisation to examine where mitigation efforts are focussed, potentially identifying areas covered by multiple interventions while others remain exposed.

Within DI, mitigation against the risk of intelligence failure is strongest in areas where interventions can be converted into standard procedures. Where these are automated or incorporated into systems, there is less scope for them to be accidentally or deliberately circumvented, but most are still reliant on humans checking they have been adhered to. This may be checks by the analysts themselves, supervisors, or at an organisational level. The SAS-114 technical team research has been useful in this area for identifying opportunities for improved or new practices that include a greater degree of automation or compulsion.

Mitigation against intelligence failures becomes harder when addressing risks arising from activities that are less easily standardised, due to higher levels of variance and therefore uncertainty. This is most apparent when dealing with human thought processes and decision making, and therefore mainly affects the risks related to judgement. The SAS-114 technical team research, along with stakeholder discussions, has been most valuable in this area, as it has provoked thought and reflection on why DI uses certain practices, providing both evidence and challenge. Further engagement and research in this area would build on this progress, enabling DI to continue to develop robust analytic tradecraft to support its critical role in decision making in support of UK Defence.

2.10 REFERENCES

[1] Jousselme, A. (2016). The risk game: Formalisation and analysis of reasoning profiles. *NATO STO Meeting Proceedings*, Annex K (PUB REF STO-MP-SAS-114-PPF). Brussels, Belgium: NATO STO.

[2] Bex, F. (2016). Computational scenarios and arguments: An AI approach to structured analytic techniques. *NATO STO Meeting Proceedings*, Annex F (PUB REF STO-MP-SAS-114-PPF). Brussels, Belgium: NATO STO.

[3] Nelson, J. (2017). Assessment of information utility: Tutorial and discussion. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, La Spezia, Italy.

[4] Kajdasz, J. (2017). Assessing effectiveness of psychological operations. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[5] Kerbel, J. (2017). Why the US IC struggles with uncertainty. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[6] Isaksen, B. (2017). Uncertainty handling in estimative intelligence: Problems and requirements from both analyst and consumer perspectives. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[7] McHenry, J. (2017). Comparative evaluation of the forecast accuracy of analysis products and a prediction market. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[8] Siegel, A. (2017). Humans in the loop: Examples of prediction markets' future role in guiding decisions. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[9] Timms, M. (2016). Measures of information usefulness in target classification. *NATO STO Meeting Proceedings*, Annex J (PUB REF STO-MP-SAS-114-PPF). Brussels, Belgium: NATO STO.

[10] Nelson, J. (2016). Optimal experimental design theory, asymmetric cost structures, and the value of information. *NATO STO Meeting Proceedings*, Annex I (PUB REF STO-MP-SAS-114-PPF). Brussels, Belgium: NATO STO.

[11] Jousselme, A. (2016). Update of the SAS-114 analysis working group on defence and security standards for uncertainty: Our collections. *NATO STO Meeting Proceedings*, Annex C (PUB REF STO-MP-SAS-114-PPF). Brussels, Belgium: NATO STO.

[12] Clausen Mork, J. (2016). Information gain and approaching true belief. *NATO STO Meeting Proceedings*, Annex H (PUB REF STO-MP-SAS-114-PPF). Brussels, Belgium: NATO STO.

[13] Dhami, M. (2016). Report on SAS-114 experiment on analysis of competing hypotheses. *NATO STO Meeting Proceedings*, Annex G (PUB REF STO-MP-SAS-114-PPF). Brussels, Belgium: NATO STO.

[14] Mandel, D. (2016). Report on SAS-114 "SRICCLE" experiment: Effect of source reliability, information credibility, and classification level on analysts' uncertainty about information accuracy. *NATO STO Meeting Proceedings*, Annex D (PUB REF STO-MP-SAS-114-PPF). Brussels, Belgium: NATO STO.

[15] Rein, K. (2016). I believe it's possible it might be: Using lexical clues to generate evidence weights. *NATO STO Meeting Proceedings*, Annex A (PUB REF STO-MP-SAS-114-PPF). Brussels, Belgium: NATO STO.

[16] Jousselme, A., and de Rosa, F. (2017). CMRE TTX: Report and discussion. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, La Spezia, Italy.

[17] Rein, K. (2017). SAS-114 and IST-132: Combining sensor and HUMINT information. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, La Spezia, Italy.

[18] Benjamin, D. (2017). The role of type and source of uncertainty from multiple climate projections. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[19] Nelson, J. (2017). Using automatic techniques to design experiments to learn about the psychology of information. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[20] de Rosa, F. (2018). The reliability game: Preliminary outcomes and future work. Paper presented by VTC at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Madrid, Spain.

[21] Jousselme, A. (2018). Conflict measurement in fusion systems. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Madrid, Spain.

[22] Claver, A. (2016). The predictive value of reporting. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Copenhagen, Denmark.

[23] Van de Meeburg, H. (2016). Netherlands' Devil's Advocate. *NATO STO Meeting Proceedings*, Annex E (PUB REF STO-MP-SAS-114-PPF). Brussels, Belgium: NATO STO.

[24] Dhami, M. (2017). SAT research under SAS-114: Overview and way ahead. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, La Spezia, Italy.

[25] Claver, A. (2017). Netherlands' Defence intelligence update. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, La Spezia, Italy.

[26] Niall, K. (2017). The true meets the possible: Deductive logics for practical reasoning. Paper presented by VTC at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, La Spezia, Italy.

[27] Dhami, M. (2017). Testing the effectiveness of scenario generation techniques. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[28] Clausen Mork, J. (2017). Intelligence analysis tasks and the outlook for evidence-based methodology choices. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[29] Rieber, S. (2017). IARPA's CREATE Program: Crowdsourcing evidence, argumentation, thinking and evaluation. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[30] de Werd, P. (2017). Analysis by contrasting narratives: Identifying and analysing the most relevant truths. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[31] Twardy, C. (2017). Frequency formats for complex probability puzzles: A live exercise. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[32] Shaw, R. (2017). By what method(s) should intelligence organisations assess threats and hazards? Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[33] Budescu, D. (2017). Using relative judgments to estimate subjective distributions. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[34] Dhami, M. (2018). Intuition, deliberation and quasi-rationality in intelligence analysis. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Madrid, Spain.

[35] Nelson, J. (2018). Information formats for presenting probabilistic information: Usefulness for search and belief updating, and implications for the IC. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Madrid, Spain.

[36] Mandel, D. (2018). Probabilistic hypothesis testing with correlated evidence: Tests of ACH and Bayesian methods. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Madrid, Spain.

[37] Græsholm, E. (2016). Precautionary intelligence: Communicating uncertainty through prediction. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Copenhagen, Denmark.

[38] Kajdasz, J. (2016). Interpretation of NATO standards by non-native English speakers. *NATO STO Meeting Proceedings*, Annex B (PUB REF STO-MP-SAS-114-PPF). Brussels, Belgium: NATO STO.

[39] Arcos, R. (2017). Communicating intelligence in a digital era. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, La Spezia, Italy.

[40] Arcos, R. (2017). Communication of risk and uncertainty by commercial open source intelligence providers. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, La Spezia, Italy.

[41] Proctor, J. (2017). Challenges in modernising and standardising expressions of uncertainty in analysis. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[42] Shaw, R. (2017). Canadian Forces Intelligence Command (CFINTCOM) approach to communicating uncertainty in intelligence. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[43] Van de Meeburg, H. (2017). Devil's advocacy and quality assurance in the Netherlands. Paper presented at the Winter Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Washington DC.

[44] Shaw, R. (2018). Developments in developing a standard for communicating event probabilities and analytic confidence at CFINTCOM. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Madrid, Spain.

[45] Arcos, R. (2018). Assessment and communication of uncertainty in EU INTCEN: A post-mortem. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Madrid, Spain.

[46] Mandel, D. (2018). Internalisation and perceived organisational compliance by Canadian intelligence professionals of US ICD-203 standards. Paper presented at the Spring Meeting of the NATO SAS-114 on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making, Madrid, Spain.



Chapter 3 – DEVIL'S ADVOCACY WITHIN DUTCH DEFENCE: IMPROVING INTELLIGENCE SUPPORT TO DECISION MAKING

A. Claver and H.M. van de Meeberg
Ministry of Defence
THE NETHERLANDS

3.1 INTRODUCTION

There is a large body of research on decision making [1], [2], [3], covering, amongst others, analytic techniques contributing to the decision-making process [4], [5], [6]. This chapter will not discuss the multitude of methods and/or techniques available, that can be used to assist in decision making and can provide checks and balances for the decision-making process. Instead, the focus will be on a single method, i.e., Devil's Advocacy and its application in a specific field of activity: intelligence. It will show how the use of this concept can contribute to making better intelligence assessments and help decision makers in setting priorities and allocating resources more effectively as exemplified by the Dutch Defence Intelligence and Security Service (DISS) since 2008.[1]

Devil's Advocacy within the DISS – inspired by and partially modelled after the Israeli experience – came into being some years after the intelligence debacle concerning Saddam Hussein's alleged weapons of mass destruction (2003). Fuelled by the highly critical reports on the failures of particularly British and American intelligence agencies it was felt within the Dutch Ministry of Defence that stricter quality control of analytical products of the DISS was of vital importance, recognizing the fact that Dutch support for the Iraq War had been decided on the basis of questionable intelligence and one-sided arguments. This led to the introduction of a small, separate Devil's Advocate team in 2008 within the DISS. This chapter describes and analyzes the functioning of the Dutch Devil's Advocate and argues for the imperative of criticism [7].

In this chapter, the concept and functioning of Devil's Advocacy is highlighted first, whereby reference is made to the Israeli example that has been in existence for almost half a century. This will provide a framework for describing and analyzing the activities of the Dutch Devil's Advocate since 2008. This is followed by an assessment of the Devil's Advocate's effects on the intelligence processes and decision making. The final part addresses the applicability of Devil's Advocacy in more general terms and offers some tentative thoughts on the imperative of Devil's Advocacy in support of governance in the new era of cyber and big data.

3.2 ENTER THE DEVIL'S ADVOCATE: CONCEPT AND METHODOLOGY

The origins of the Devil's Advocate are centuries old and lie within the Roman Catholic Church, where the *Advocatus Diaboli* – with the formal title of *Promotor Fidei*, i.e., Promoter of the Faith – fulfilled the distinct purpose of challenging the purported virtues and miracles of nominees for sainthood. The methodological approach used was to argue against the nominee's case on the basis of the same sources provided in his or her favour and to uncover any misrepresentation of the evidence favouring canonization during that process. The official position of the Devil's Advocate enabled Rome to take control of a haphazard and decentralized process of saints in the making that was spinning out of control, and was considered to be a threat to the centralized authority of the Catholic faith. The mitigating effects of this decision can be illustrated by observing what happened after the decision of Pope John Paul II to basically eliminate the Devil's Advocate

[1] The *Militaire Inlichtingen- en Veiligheidsdienst* (MIVD; English acronym DISS) is an all-source service responsible for defence-related intelligence, counter-intelligence and security issues, and comprising intelligence, counter-intelligence and security personnel as well as OSINT, IMINT/GEOINT, HUMINT, and SIGINT capabilities.

office in 1983 after nearly four hundred years. Subsequently, until his death in 2005 there were more beatifications (1,338) and canonizations (482) declared than his 263 predecessors combined managed in over almost two thousand years (see Ref. [8], ix-xii).

3.2.1 Concept

In common speech the term Devil's Advocate applies to anyone with a dissenting view that takes a contrary position (for the sake of argument alone, either authentic or non-authentic). In this understanding of the term the contrarian dissenting view takes central position. The famous English philosopher and economist John Stuart Mill has been an early advocate for the importance of dissent. In the middle of the nineteenth century he argued that people tend to deceive themselves by taking the veracity of unexamined convictions for granted. In his opinion such conformity would lead to passive acceptance of the world as it is. Ultimately, succumbing to this kind of behavior would result in what he referred to as "tyranny of the majority." Mill therefore advocated for people to be exposed to as wide a range of opinions as possible and urged for no idea or practice to remain unchallenged [9].

Devil's Advocacy nowadays is considered an established contrarian technique within the growing research body of structured analytic techniques. It is defined as challenging a strongly held view or consensus by building the best possible case for an alternative explanation (see Refs. [4]; [6], pp. 17-18). In doing so, Devil's Advocacy aims to serve as a check on a dominant mindset that considers only one (or a limited number of) aspect(s) of a given issue. This so-called groupthink will dismiss contradictory evidence or fail to give it its proper weight or consideration as a result of which the quality of the assessment suffers (see Refs. [1]; [5], pp. 217-223; [10], p. 29).

Within academia Devil's Advocacy is recognized as a method of critical thinking that can contribute to problem-solving and decision-making processes. This is not to say that the concept of Devil's Advocacy has not received any criticism [11], [12]. Instances of the negative use of the concept have been recorded, especially when the dissenting view was offered by an individual or a team specifically tasked to do so. According to Charlan Nemeth the value of dissent is unmistakable, but she emphasizes the importance of authenticity. Her research on minority influence has shown that authentic dissent – even of individuals – can influence decision-making processes [13]. It has also provided evidence that dissent – even when it is wrong – stimulates divergent thinking and the consideration of alternatives that ultimately serves the quality of decisions [14].

3.2.2 Methodology

The Devil's Advocate essential role is to test the validity of propositions by seeking to prove the opposite of whatever the challenged view holds. This aim is achieved not by asking whether an analysis is right or wrong, but whether it is self-consistent. From there it is a small step from testing the self-consistency of assessments to testing or confirming its findings (see Refs. [7], p. 66; [10], p. 29). Application of the concept is of added value in two additional ways. First, it adds to the robustness of the assessment by considering a topic from different angles. Second, it helps tackle the neglected, but fundamental problem of the so-called 'Alpha and Beta chance'.[2]

The Alpha chance is the chance that you incorrectly conclude that there is a significant relationship between phenomena. The Beta chance is the chance that you fail to discover a weak relationship between phenomena. In scientific analysis the alpha chance is placed low: usually at 5%. In other words, evidence of the relationship between described phenomena is only accepted if it has been observed in at least 95 out of 100 instances. The beta chance is usually positioned much higher: between 20 – 90 %. Or, to put it differently, it is accepted that weak relationships – that actually exist – are missed (see Refs. [10], pp. 66-68; [15], pp. 14-16, p. 37).

[2] English and American readers will probably be more familiar with the expression Type I and Type II instead of Alpha and Beta. The terminology is, however, interchangeable and expresses the same elements when used.

These percentages, however, are not acceptable when it comes to risk calculations and explanations in cases of e.g., war (threats) or terrorist activities. Identification of existing relationships is paramount from the perspective of intelligence and security, and timely warning. In these instances, the beta chance should be put lower (in order to not miss a weak, but vital relationship), and the alpha chance higher (making a relationship significant sooner). Robust research through the application of different methods and techniques, like Devil's Advocacy, Red Teaming, and Analysis of Competing Hypotheses (ACH) appears to minimize the margin of error when it comes to the Alpha Chance [4], while maximizing the chance of discovering significant relationships (Beta Chance) (see Refs. [10], pp. 67-68; [16]).

3.3 THE DUTCH EXPERIENCE

What activities have been undertaken by the Dutch Devil's Advocate after its establishment and what results did they bring? What are the findings of the past decade and how can they be assessed? And, does the Dutch experience with Devil's Advocacy support the above assertions? Though scarce, the available sources seem to allow a tentative answer to the questions posed above. In order to put the Dutch findings into better perspective first some attention is given to the Israeli efforts in this field, followed by a more elaborate and detailed description of the Dutch case.

3.3.1 Devil's Advocacy in Israel[3]

Despite the availability of a wealth of (intelligence) information before the outbreak of hostilities, the Yom Kippur War showed poor comprehension on all executive levels – individual, organizational and governmental – as a result of which decision making suffered badly [17], [18], [19]. After conclusion of the war the so-called Agranat Commission[4] reported major shortcomings in the assessments of military intelligence regarding Arab intentions and capabilities. The shock of war encouraged a variety of reform activities (see Refs. [20], p. 3; [21], pp. 210-213). One of these reforms was the creation within Israeli military intelligence[5] of the Mahleket Bakara – i.e., Department of Control, also known as Ipcha Mistabra, which means "The opposite is possible." [23]. According to Refs. [8], p. 250, and [24], Mahleket Bakara is sometimes also translated as "Research Unit" or "Internal Audit Unit".

The novel concept of Devil's Advocacy was introduced and designed to act as a safety valve (amongst others to counter phenomena like groupthink and tunnel vision) and contribute to the production of timely, relevant and correct intelligence assessments. The Devil's Advocate reports directly to the head of military Intelligence and, since 1996, to the Prime Minister and Minister of Defence as well.[6] The new office also aimed to further a culture of openness within Aman where individuals are expected to voice dissenting opinions. Aman's organizational slogan bears witness to this openness: "Freedom of opinion, discipline in action." (see Refs. [8], p. 63; [20], p. 4; [24]; [25]). Most sources argue that the existence of a Devil's Advocate, as an institutional-level safeguard against groupthink, has instilled an atmosphere of accountability within the analytical process. Analysts today have to stand behind their analysis and be prepared to deal with critique (see Refs. [20], p. 4; [24]).

[3] With regard to this topic, the authors have spoken with three officers of the Israeli Defence Forces (IDF) (15 December 2009, 5 December 2017, 13 June 2018) and Professor Shlomo Shpiro, Bar-Ilan University, Tel Aviv (28 October 2016).

[4] The Agranat commission – named after its chairman, Chief Justice of the Supreme Court Dr. Shimon Agranat – was installed on 21 November 1973 and asked to examine: "The intelligence available from before the war on the intentions of Syria and Egypt; the analysis of the intelligence by the authorized civilian and military units; the general preparedness of the IDF to fight, especially on the date of October 5th 1973, the day prior to the outbreak of the war." See Refs. [19], pp. 499-516; [21], pp. x-xvi; [22].

[5] The military intelligence service (called Aman) is Israel's only agency capable of providing a holistic intelligence picture (see Ref. [20], 3-4). Its primary task to date is intelligence analysis, whereas the other intelligence agencies (Mossad and the internal security service, known by its acronym Shabak or abbreviation Shin Bet) are basically operational organizations responsible for intelligence collection, covert operations and counter-terrorism.

[6] Personal conversation with Professor Shlomo Shpiro, 28 October 2016.

Israel's Devil's Advocate has come to be accepted as a legitimate feature. No (publicly voiced) calls have been made to abolish it although (internal) critique has never been far removed (see Refs. [7], pp. 66-67; [21], pp. 213-216; [24]). The contradictory stance of the Devil's Advocate has served the purpose of acting as a check on organizational routine within military intelligence, the Defence ministry at large and on occasion within the Prime Minister's office [24]. Unlike 1973, decision makers can now be accessed directly and provided with additional (contrarian) points of view, thus strengthening their information base and making the decision makers' choices more robust.

3.3.2 Devil's Advocacy in the Netherlands[7]

Within the Netherlands the controversy since the start of the Iraq war in 2003, regarding the intentions of Saddam Hussein's regime and its (non-existing) weapons of mass destruction, called into question the validity of intelligence assessments and the effectiveness of existing quality control procedures within the intelligence services.[8] Similar discussions took place elsewhere leading to (parliamentary) inquiries in various countries. Early July 2004 the British and American inquiries concerning the functioning of the intelligence services in the run-up to the Iraq war were published. Being highly critical, the findings of the Butler and Roberts committee caused a stir in both countries. This reverberated within the highest Dutch political and administrative levels. The DISS leadership was subsequently tasked to see whether the introduction of a Devil's Advocate would be a useful addition to existing analytical procedures to prevent a similar outcome from occurring in the future.

A DISS report on the subject, completed in June 2005, included a review of relevant literature, interviews within the organization itself as well as a review of the Israeli and British approach to analytical quality control. The report concluded, that the introduction of a Devil's Advocate (*Advocaat van de Duivel* or *Duivelsadvocaat* in Dutch) as an independent organizational element – analogous to the Israeli example – could not be recommended. Three main reasons were given for this outcome:

- The perceived lack of internal acceptance of such a function constituting too much of a top-down approach, as well as the lack of a candidate with suitable seniority and the right professional qualifications.
- The limited processing capacity of a Devil's Advocate (office), which would make it impossible to cover all of the analytical output.
- The insurmountable problems related to the selection of products to investigate.

The report did not, however, downplay the need for better overall quality assurance. The need for improved work processes and higher-quality products was strongly emphasized. This entailed, for instance, more professional training and instruction of individual analysts, better registration and uniform implementation of analytical work processes, the adoption and implementation of selected research methods and techniques, and the structural use of the peer reviewing technique. Interestingly, the possibility of introducing distinct product reviewers within each analytical team was also suggested, as well as the appointment of a Professional Head of Intelligence Analysis – analogous to the British example – responsible for the content of professional training and the analytical competencies required of all analysts. This proposal certainly touched upon elements as practised within Devil's Advocacy. Following up on this proposal would, however, come at considerable cost, since it was estimated that it would entail the hiring of ten to twelve additional senior staff members. The effort and budget required to implement all these recommendations was substantial, and considered not feasible at the time.

[7] The following section is primarily based upon interviews with the Dutch Devil's Advocate himself, and his team, conducted in August 2018. Additional information can be found in DISS annual reports (see Ref. [26]: 2008, p. 57; 2010, p. 61; 2011, p. 49; 2012, pp. 9, 13, 57; 2013, pp. 9, 16; 2015, p. 9. See also Ref. [27], pp. 10-11).

[8] A Dutch Iraq-inquiry in 2010 would ultimately judge positively regarding the quality of DISS reporting in the run-up to the war.

In 2006, however, the internal and external landscape changed in a number of ways for the DISS. Firstly, the defence intelligence and security community as a whole was investigated by a special independent committee, named after its chairman C.W.M. Dessens. The committee Dessens came to the following conclusion with regard to the DISS (see Ref. [28], p. 45):

> *"The quality of the DISS-products requires constant attention. Therefore, it is necessary that the DISS develops a uniform quality control system (kwaliteitszorgsysteem) on the basis of which the quality of the DISS-products is structurally secured and monitored. […] In addition, it needs to be seen whether it is desirable to realize additional quality control with regard to certain important products."*[9]

Secondly, an increase in the budget of the service allowed the intake of additional personnel making it more feasible to address the issue of quality assurance. Thirdly, the new leadership of the service strongly embraced the goal of better quality control in both intelligence collection and analysis, while expressing at the same time the need for an experienced senior special advisor to assist the DISS Director and his management team in policy as well as operational decision making.

These three factors led to the introduction of a Devil's Advocate office within the DISS as of 1 January 2008 consisting of a Devil's Advocate, concurrently senior policy advisor, and a small team of senior staff members with long careers in intelligence collection and/or analysis. Note should be taken of the fact that the Devil's Advocate in this Dutch model was designed for quality assurance purposes in the broadest sense, covering basically the entire intelligence cycle, i.e., starting from the intelligence requirement to the distribution of the intelligence product (as well as the customers' feedback).

The activities of the Devil's Advocate addressed, on the one hand, an urgent need of policy and decision makers at the highest executive and administrative level for intelligence assessments of consistent and high quality, based not only on effective intelligence collection, but also on the correct application of a variety of structured analytical techniques. These assessments should be high-quality elements upon which political decisions could (at least partially) be based. The Devil's Advocate office being responsible for contrarian analysis and critical reviews of collection efforts and analytical products was set up to play a vital role in this field. On the other hand, The Devil's Advocate's own expertise and the know-how resulting from the reviews and enquiries by the other senior staff members of the Devil's Advocate into the functioning of the entire intelligence cycle were to enable the Devil's Advocate office to do much more than that. The Devil's Advocate was supposed to provide the DISS leadership with an answer to a much broader question: Do we do the right things in carrying out our legal intelligence tasks and do we do them effectively?

The design of the Devil's Advocate team showed the following distinguishing characteristics:

1) It was recognized from the start that successful functioning of a Devil's Advocate team required (or more accurately: presupposed) a high level of professionalism within the organization. The team could only function effectively on the condition that a self-critical mentality was part and parcel of every day working conditions in the service.
2) It was realized that to be successful the Devil's Advocate activities required a support base within the service. However, trust and respect would need to be earned, since the initiative was started top-down instead of bottom-up. To build a support base, transparency was deemed of great value.

[9] A few years after Dessens, in 2010, the Davids committee – the belated Dutch parliamentary inquiry into the Iraq war which, incidentally, toppled the coalition government of Prime Minister Balkenende after the publication of its report – would elaborate on this element of quality control by emphasizing the importance for the government to make its own independent consideration (*eigenstandige afweging*); i.e., not having to rely (completely) on information and/or intelligence of other parties. In other words, the civilian and military intelligence and security service should be able to collect and/or validate sources themselves and supply the government with (at least partially) independent assessments (see Ref. [29], pp. 117, 273, 507). Self-dependence (*eigenstandigheid*) was subsequently taken up by the Devil's Advocate as an element of major concern in its investigations.

For this reason, an annual activity plan was formulated and discussed beforehand with the various production and collection departments, definitions, criteria and formats for product reviews were drafted and made available, and all Devil's Advocate reports and reviews could be accessed and read by every employee interested. The Devil's Advocate, furthermore, made sure to introduce himself to all employees through an internally published interview and frequent presentations (for new and old personnel alike).

3) It was understood that to execute his tasks with any chance of success the Devil's Advocate needed to have unlimited access to all available relevant information circulating within the organization: i.e., raw intelligence, work processes, semi-products and final products. Internal procedures were set up to accommodate this.

4) Finally, it was deemed crucial that the Devil's Advocate would be completely independent from the main production and collection departments. Since the Devil's Advocate himself played a second role as senior policy advisor, the entire Devil's Advocate Office team was made accountable to the Director only. The Office could therefore not be seen as part of the line organization nor be made (partially) accountable for disseminated products. As a result, the Devil's Advocate focus was on *ex post* reviews of analytical products and on *ex ante* advice, and enquiries into collection, processing and distribution activities.

To contribute to the quality improvement of DISS intelligence products and the processes underlying these products the Devil's Advocate focused on the triad: collection, information processing and analysis, and production. The collection efforts, the organization's information processing mechanism (including work flow management), and the analytical production (i.e., content of the output, dissemination, and client feedback) by the DISS has been subject of critical research by the Devil's Advocate team ever since.

In its decade of existence (2008 – 2018) the activities of the Devil's Advocate have seen two distinctive phases:

1) ***Performance assessment: focus on the "How" (2008 – 2011).*** This first phase emphasized research activities, assessment and advice on quality improvement of individual intelligence products, collection processes and related analytical skills with the purpose to prevent and counteract (analytical) tunnel vision, and groupthink. The main tools used to achieve this were critical surveys, frequent product reviews (according to a fixed format and transparent quality criteria), the organization of training courses and seminars in order to 'improve the analyst'. At the same time the Devil's Advocate assisted in setting up a certified Master Program on Intelligence within the national Dutch Defence Academy. The training in and use of structured analytic techniques by analysts was encouraged. Seminars with outside experts on international topics were organized to stimulate "new thinking" among the analysts. These (ongoing) activities were executed ex post in order not to interfere with current production and intelligence research assignments. The focus was always to link Devil's Advocate assessments with advice on how to realize improvements in future products and collection efforts.)

2) ***Performance assessment: focus on the "What and Why" (2012 – 2018).*** In the second phase the Devil's Advocate team focused in more generic terms on the earlier-mentioned questions: do we do the right tasks, and we do them effectively. For that purpose, the activities of the Devil's Advocate team included the design and implementation of a qualitative system, capable of dealing with the problem of scarcity of resources. This so-called 'Weighing and Prioritizing' (W&P) system is a tool to assist in allocating (scarce) resources to areas of intelligence interest and research assignments, in consultation with the "customers" of the products of the DISS. The W&P tool was supplemented by assessments of the relevance of collection assets and by the input of a "matrix", a management tool to quantify and bring ambitions of the DISS and customers' wishes and intelligence requirements in balance with available scarce resources. An important addition to the W&P system was the customer

feedback research organized by the Devil's Advocate, which asked the important question whether products and services of the DISS had met the jointly set requirements? This feedback closed the well-known intelligence cycle: i.e., the collection, processing, analysis, and product dissemination loop. Last but not least, the Devil's Advocate team in this phase made detailed assessments of terminology and time indicators used in the analytical products. Recently, the team also looked into the predictive value of analytical judgements by means of calibrated feedback. This ongoing investigation aims to verify the accuracy of the forecasts made in order to assess the quality and usefulness of analytical output, and support much needed outcome accountability.

What were the effects of the work of the Devil's Advocate? Did the concept work out as planned? The overall balance is positive. Over the years the Devil's Advocate and his team have become widely accepted within the DISS. Evaluation of the project and its relevance after nearly three years led to formalization of the Devil's Advocate concept within the DISS as of 1 January 2011. Quality improvements in analytical products and greater effectiveness in collection, more effective prioritization in the use of collection assets and of research assignments as well as greater sensitivity within the entire service to customers' priorities were indeed achieved.

However, the effectiveness of a Devil's Advocate's top-down approach has its limits. Devil's Advocacy cannot rely on formats, new management tools (like W&P, and the "matrix" instrument) and built-in routines alone. It is very much a live concept and should be treated as such. The experience of the much older Israeli Devil's Advocate office over the years reflects this as well (Lieut. Col. S. and V.I. 2017). Organizational and analytical relapses have indicated that progress is always fragile, requiring continued 'investment' on all levels. Assessments and reviews need to be repeated to prevent the quality of products and their underlying processes from sliding. Continued and open support of line management across the board remains essential, as well as the strongest backing from the highest decision-making level. Because, in agreement with Ref. [8], p. 263, and several experts in the field, lack of access and lack of support in combination with organizational inertia/opposition (passive and/or active) ultimately render any Devil's Advocate toothless and ineffective.[10]

3.4 DEVIL'S ADVOCACY ON THE HORIZON?

What might be in store for the future? The overwhelming wealth of data nowadays can only be handled 'industrially' through 'data mining', i.e., the automated processing by means of software algorithms. The DA's unrestricted access to data mentioned before now equals a data deluge that will paralyze any individual (or group of individuals) trying to understand what these bits and bytes actually have to tell.

It follows that the application of a contradictory perspective requires DA to refocus on the research of inputs (i.e., processes); not merely outputs (i.e., products). Hard copy products will still be reviewed in the traditional sense. No doubt decision makers will keep receiving paper, but a Devil's Advocate operating in the (cyber) future cannot base an assessment on physical output alone. This shift in focus in itself does not constitute a problem, as evidenced by the Israeli and Dutch Devil's Advocate, who consider it mandatory to investigate processes and products when forming an opinion [23].[11] In other words, Devil's Advocacy will have to understand software and hardware processes and most likely will have to turn the attention to algorithm reviews.

These reviews can serve as a check on the important issue of correlation vs. causality, hiding itself within each of the algorithms in use. Attaching meaning to data by correlating them touches upon a crucial element of (big) data processing within digital space [30]. Often correlation is mistaken for causation, although a

[10] Personal conversation with IDF officers, 15 December 2009 and 5 December 2017, and the Dutch Devil's Advocate, 23 August 2018.

[11] Personal conversation with IDF officer, 13 June 2018, and the Dutch Devil's Advocate, 23 August 2018.

correlation merely quantifies the statistical relationship between two data points. When one data point changes, the other is likely to change in case of a strong correlation. This change is less likely to occur in case of a weak correlation. However, even strong correlations might occur because of coincidence. After all, correlation "only" implies probability and an analysis based on statistical probabilities will, by definition, produce false positives (e.g., criminalizing innocent people) and false negatives (e.g., allowing security risks to go unnoticed).

Analysis today is driven more and more by the (over)abundance of data, i.e., so-called big data. The main challenge, however, is how to cope with the variety, messiness, and uncertainty of the compiled data set. People have to bear in mind that much of what is collected does not have a specific question in mind, or is the (unintended) by-product of another activity. Data algorithms seek to find interesting links and identify relevant patterns. Although they might provide unexpected insights, they also run the risk of elevating correlations to causations. Or, increase the possibility of making one or more false discoveries, i.e., the Alpha chance or Type I error mentioned earlier, when crunching huge amounts of data. Big data seems to elevate the risk of that sort of error. A Devil's Advocate offering a contradictory perspective can challenge these algorithms and prevent them from being taken at face value (see Refs. [5], pp. 217-218; [6], pp. 1-2). Notwithstanding the (future) possibilities of big data applications there will always remain a gap in data collection and a resulting lack of situational awareness, which is of critical importance for decision makers. To become and stay relevant in the foreseeable future, (cyber) analysis has to bridge the strategic, operational, and tactical domains in order to provide insight at each level of analysis [31]. Devil's Advocacy can play a role here in addressing the need to overcome the tactical-technical focus and reviewing processes and products on their ability to offer decision makers cyber analyses they can actually use.

Devil's Advocacy certainly does not represent the holy grail of problem solving. It is but one of several tools analysts and decision makers can use in (cyber) considerations. See, for example, the possibility of cyber red teaming as described in Ref. [32]. Any tool is obviously limited in its usefulness and applicability. People instinctively resist a Devil's Advocate opinion precisely because it purposefully threatens an argument that might have taken much time and effort to arrive at (see Ref. [5], p. 219). Likewise, application of the Devil's Advocate technique itself consumes a disproportionate amount of time and effort. For these reasons it needs to be practised sparingly (see Refs. [8], p. 263; [20], pp. 3-4).[12]

Finally, when investigating intelligence processes and products account should be taken of the thin line between offering a critical review and contradictory perspective, and the concept of oversight. Devil's Advocacy is about internal accountability, quality assurance and expediency (both issues of efficiency and effectiveness), whereas an oversight mechanism deals with external accountability and legitimacy. The two concepts should be seen as different mechanisms serving similar purposes. They are not necessarily mutually exclusive and, in fact, can complement each other (see Ref. [33], pp. 1426-1438). External oversight teams and internal Devil's Advocate teams, review teams, inspection teams, or auditing teams have much to learn from each other. For this reason, a network of partners assigned to and/or interested in cyber quality assurance might well be worth considering.

3.5 CONCLUSION

In today's unsettled world the (in)experienced cyber decision maker has an obvious need for a differing perspective. Devil's Advocacy is very much a live concept, capable of adapting itself to cope with new challenges as shown by the Dutch and Israeli examples. Devil's Advocacy in the Dutch model is about internal accountability, transparency and quality assurance in terms of both effectiveness and efficiency in all phases of the intelligence cycle. This will remain so in the coming years. At the same time, with the help of a Devil's Advocate decision makers can be offered another 'convincing' perspective to raise awareness of new issues at stake and try to prevent the cognitive pitfalls present. Pitfalls that can be ill afforded. The imperative

[12] Personal conversation with Professor Shlomo Shpiro, 28 October 2016.

of criticism remains and Devil's Advocacy serves the purpose of testing the validity of propositions put forward. By offering a different perspective, battling groupthink, assigning Alpha and Beta chance problems their proper place, and reducing the information gap, the activities of the Devil's Advocate can also in the new world of cyber help to improve the robustness of assessments and, by definition, the informed position of decision making.

3.6 REFERENCES

[1] Coulthedart, S. (2016). Why do analysts use structured analytic techniques? An in-depth study of an American intelligence agency. *Intelligence and National Security* 31:933-48.

[2] Akoto, W. (2014). Paradigms of foreign policy and political decision making. A critical review of three seminal works. Accessed 27 November 2016. https://www.researchgate.net/publication/261707971_Paradigms_of_Foreign_Policy_and_Political_Decision_Making_A_Critical_Review_of_Three_Seminal_Works.

[3] Buchanan, L., and O'Connell, A. (2006). A brief history of decision making. *Harvard Business Review* 84:32-41.

[4] Heuer, R.J., and Pherson, R.H. (2010). *Structured Analytic Techniques for Intelligence Analysis*. Washington DC: CQ Press.

[5] Jones, M.D. (1998). *The Thinker's Toolkit: 14 Powerful Techniques for Problem Solving*. New York, NY: Three Rivers Press.

[6] Central Intelligence Agency. (2009). *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis*. Prepared by the US government. Accessed 16 August 2018. https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/booksand-monographs/Tradecraft%20Primer-apr09.pdf.

[7] Shmuel, Lt. Col. (1985). The imperative of criticism. *Studies in Intelligence* 24:65-70.

[8] Zenko, M. (2015). *Red Team: How to Succeed by Thinking like the Enemy*. New York, NY: Basic Books.

[9] The Economist. (2018). Against the tyranny of the majority: John Stuart Mill's warning still resonates today. 4 August 2018. Retrieved from https://www.economist.com/schools-brief/2018/08/04/against-the-tyranny-of-the-majority.

[10] de Valk, G. (2005). Dutch intelligence – towards a qualitative framework for analysis: With case studies on the Shipping Research Bureau and the National Security Service (BVD). Ph.D. thesis, University of Groningen, Groningen, the Netherlands.

[11] Stack, K.P. (1997). A negative view of competitive analysis. *International Journal of Intelligence and CounterIntelligence* 10:456-464.

[12] Mitchell, G.R. (2006). Team B intelligence coups. *The Quarterly Journal of Speech* 92:144-173.

[13] Nemeth, C.J. (2010). Minority influence theory. IRLE Working Paper No. 218-10. Accessed 30 July 2018. http://irle.berkeley.edu/workingpapers/218-10.pdf.

[14] Nemeth, C.J. (2018). *In Defence of Troublemakers. The Power of Dissent in Life and Business*. New York, NY: Basic Books.

[15] de Zoete, T.S. (2013). Alpha proof, Beta check? De mogelijkheid en wenselijkheid van een organisatorische scheiding tussen α- en β-kans gerichte activiteiten binnen politie Amsterdam-Amstelland op het terrein van politiek gemotiveerd geweld, gezien de ethische aspecten die bij een dergelijke scheiding tussen inlichtingenwerk en opsporing aanwezig zijn. Master's thesis, Leiden University, Leiden, the Netherlands.

[16] Goldbach, O., and de Valk, G. (2016). To explore the unknown. Towards a methodology of not to miss a threat (Rumsfeld matrix). Presentation at the conference Witness to Change: Intelligence Analysis in a Changing Environment, NISA 25th Anniversary Conference 1991 – 2016', The Hague, 27 – 28 October 2016.

[17] Bar-Joseph, U. (1995). The wealth of information and the poverty of comprehension: Israel's intelligence failure of 1973 revisited. *Intelligence and National Security* 10:229-240.

[18] Riedel, B. (2017). *Enigma: The Anatomy of Israel's Intelligence Failure Almost 45 Years Ago*. Washington DC: Brookings Institution.

[19] Rabinovich, A. (2004). *The Yom Kippur War. The Epic Encounter That Transformed the Middle East*. New York, NY: Schocken Books.

[20] Kuperwasser, Y. (2007). *Lessons from Israel's Intelligence Reforms*. Washington DC: Brookings Institution.

[21] Shalev, A. (2010). *Israel's Intelligence Assessment before the Yom Kippur War: Disentangling Deception and Distraction*. Portland, OR: Sussex Academic Press.

[22] Agranat Commission. (2016). https://www.knesset.gov.il/lexicon/eng/agranat_eng.htm. Accessed 13 November 2017.

[23] Lt. Col. S. and V.I. (2017). Intelligence supervision: Creating relevance in the present era. *Intelligence in Theory and in Practice: A Journal on Intelligence Methodology* 2:121-130.

[24] Shpiro, S. (2016). The devil's advocate. Controlling Israel's military intelligence analysis. Presentation at the conference Witness to Change: Intelligence Analysis in a Changing Environment, NISA 25th Anniversary Conference 1991 – 2016, The Hague, 27 – 28 October 2016.

[25] Pascovich, E. (2018). The devil's advocate in intelligence. The Israeli experience. *Intelligence and National Security.* Published online 7 May 2018, 33:854-865.

[26] *Militaire Inlichtingen- en Veiligheidsdienst Jaarverslag*. The Hague: MIVD. https://www.inlichtingendiensten.nl/organisatie/jaarverslagen-0.

[27] Ingelicht. (2008). Advocaat van de duivel in dienst. *Ingelicht*, February 2008, pp. 10-11.

[28] Dessens, C.W.M. (2006). *Inlichtingen en Veiligheid Defensie. Kwaliteit, Capaciteit en Samenwerking.* The Hague: Ministerie van Defensie.

[29] Davids, W.J.M. *et al.* (2010). *Rapport Commissie van onderzoek besluitvorming Irak*. Amsterdam, Netherlands: Boom.

[30] Claver, A. (2018). The big data paradox: Juggling data flows, transparency and secrets. *Militaire Spectator* 187:309-323.

[31] Little Limbago, A. (2014). The great divide. Closing the gap in cyber analysis. Accessed 23 April 2018. https://www.endgame.com/blog/technical-blog/great-divide-closing-gap-cyber-analysis.

[32] Brangetto, P., Emin, Ç, and Rõigas, H. (2015). Cyber red teaming: Organisational, technical, and legal implications in a military context. Tallinn, Estonia: CCDCOE. Accessed 13 November 2016. https://ccdcoe.org/sites/default/files/multimedia/pdf/Cyber_Red_Team.pdf.

[33] Mishkin, B.S. (2013). Filling the oversight gap: The case for local intelligence oversight. *New York University Law Review* 88:1414-88.



Chapter 4 – INTELLIGENCE PROFESSIONALS' VIEWS ON ANALYTIC STANDARDS AND ORGANIZATIONAL COMPLIANCE[1]

Tonya L. Hendriks and David R. Mandel
Defence Research and Development Canada
CANADA

4.1 INTRODUCTION

Deficiencies in analytic tradecraft have been cited as an enabler of recent failures among the United States (US) Intelligence Community (IC) [1], [2]. In fact, the need to improve intelligence gathering and information sharing was recognized immediately after 9/11 [3]. On December 17, 2004, the structural organization of the agencies that comprise the US IC was changed under the Intelligence Reform and Terrorism Prevention Act [4]. The IRTPA addressed the belief that weak analytic tradecraft had been an underlying cause of intelligence failures in the US by requiring the Director of National Intelligence (DNI) to establish and enforce tradecraft that meets a standard of high analytical rigor throughout the US IC [1].

On 21 June 2007, the Office of the Director of National Intelligence (ODNI) published the Intelligence Community Directive (ICD) 203, Analytic Standards [5]. ICD 203 established analytic standards (best practices) designed to improve the quality, relevance of, and confidence in the analysis and conclusions of intelligence products produced for policymakers and military commanders [6]. The 2007 version of ICD 203 consisted of five IC analytic standards and eight standards of proper analytic tradecraft [5]; a ninth analytic tradecraft standard was added in the 2015 update to ICD 203 [7].

ICD 203 includes the following five IC analytic standards:

- Objectivity;
- Independence of political consideration;
- Timeliness;
- Utilization of all available sources of intelligence information; and
- Analysis that implements and exhibits analytic tradecraft standards.

ICD 203 further includes these nine analytic tradecraft standards:

- Properly describing quality and reliability of underlying sources, data, and methodologies;
- Properly expressing and explaining uncertainties or confidence in major analytic judgments;
- Properly distinguishing between underlying intelligence information and analysts' assumptions and judgments;
- Incorporating analysis of alternatives;
- Demonstrating customer relevance and addressing implications;
- Using clear and logical argumentation;
- Explaining change to or consistency of analytic judgments;

[1] Funding support for this work provided by the Canadian Safety and Security Program Project CSSP-2016-TI-2224 (Improving Intelligence Assessment Processes with Decision Science).

- Making accurate judgments and assessments; and
- Incorporating effective visual information where appropriate.

ICD 203 aims to assist analysts from across the US IC to communicate more effectively with each other by utilizing the same standards, tools, and data as well as enabling analytic collaboration across the community [6]. As noted by Rojas [8], community-wide standards are essential for analysts to be effective in how they communicate assessments to their customers, as well as to ensure an intelligence community culture rather than a culture of independent analysts (e.g., Central Intelligence Agency, National Security Agency, Air Force analysts, etc.). In the eyes of the consumer, the adoption of common IC standards and increased analytic collaboration leads to the improvement and value enhancement of the final analytic product [9].

ICD 203 has its roots in a much longer effort to develop analytic tradecraft for the US IC. Marchio [2] provided historical evidence based on US IC declassified national intelligence assessments from 1947 through the 1990s which revealed that the US IC supported many of the elements of ICD 203 standards as early as the 1950s. For example, he noted that remarks by the Director of Central Intelligence at the Oct 20th, 1950 Intelligence Advisory indicate that the founders of the US IC were aware of the need to produce objective, relevant, and authoritative analysis that exploited all available information and acknowledged alternative views. Marchio also provided evidence to show that the early US IC recognized the value of describing quality and reliability of sources, caveating and expressing uncertainties or confidence in analytic judgments, distinguishing between underlying intelligence and analysts' assumptions, incorporating alternative analysis where appropriate, indicating relevance to US national security, using logical argumentation, highlighting change or exhibit consistency, and providing accurate judgment and assessments. Nevertheless, he argues that the US IC's commitment to developing and promoting analytic tradecraft was intermittent through the decades until intelligence oversight (e.g., congressional reviews) mandated more formal overhauls of the system. Consequently, ICD 203 has been viewed as one of the most significant accomplishments in recent IC restructuring because it formalized the standards for good analytic tradecraft, an issue that was often discussed over the history of the IC, but rarely formally documented [10].

4.1.1 The Present Research

Although ICD 203 has been implemented for several years, little is known about how IC experts view these standards and how well they think their host organizations comply with those directives. Recent research indicates that it may be difficult for assessors to reliably rate analytic product quality based on the standards unless ratings are aggregated [11].

In the present research, we aimed to examine the extent to which Canadian IC experts agreed with the directives for promoting analytic rigor captured in ICD 203. Although some Canadian intelligence professionals would be aware of ICD 203, Canada's IC is not mandated to follow ICD 203, and Canada has no national equivalent to ICD 203. Thus, it would be instructive to see how IC experts view the ICD 203 elements in cases where there is no institutional pressure to agree. Moreover, we explored the factor structure of the 13 items that were used to tap attitudinal support for the ICD 203 facets. Doing so might prove useful for conceptualizing the main components of analytic rigor as currently captured in ICD 203. We also examined the extent to which these experts judged their organizations as being in compliance with the ICD 203 directives. Because the items we used to test personal agreement and organizational compliance were matched sets, we were also able to gauge where experts perceived the largest discrepancies between their professional values and their organizations' behavior.

A distinct aim was to examine psychological correlates of attitudinal support for the professional values captured in ICD 203 and experts' judgments about organizational compliance. Specifically, we hypothesized that experts who scored high in conscientiousness and actively open-minded thinking (AOT) would show

relatively stronger support for the ICD 203 facets. Conscientious individuals tend to be achievement oriented and they value efficiency, organization, dependability and self-discipline [12]. AOT measures one's openness to evaluating information that goes against one's beliefs and to considering alternative perspectives [13], [14], [15], [16]. Individuals who score high in AOT, we anticipated, will be more likely to endorse items that reflect the importance of being fair-minded and having unbiased values that ICD 203 captures. We also hypothesized that experts who viewed their organization as meeting a relatively high level of compliance with the professional values embodied in ICD 203 will be more satisfied with their job and will show greater affective commitment to their organization. In short, we expected these individuals to be happier with their job environment. We also hypothesized that experts who perceived their organization as being in strong compliance with the professional values captured in ICD 203 will report stronger normative commitment, expressing a stronger sense of duty to the organization [17]. The hypothesis is based on the idea that we accord more commitment out of a sense of duty and respect to people and institutions we believe have acted deservingly.

4.2 METHOD

4.2.1 Participants

Participants were recruited by email with the assistance of their managers and trainers.[2] A total of 109 participants (72.2% males) completed the entire study. Participants consisted of professionals working in the Canadian IC (e.g., those attending relevant training courses, analysts, managers, etc.) who were proficient in English. The mean age was 38.4 (standard deviation [SD] = 9.40) and the mean number of years working in the intelligence community was 8.92 (SD = 7.89). The sample consisted of 50.0% civilian and 50.0% Canadian Armed Forces (CAF; note one participant did not respond). Among the military participants, the largest rank subgrouping was 'Junior Non-Commissioned Member' (NCM; 22.2%), followed by 'Junior Officer' (13.9%), 'Senior Officer' (9.3%), and 'Senior NCM' (4.6%). The highest education level groupings were 'university graduate degree' (44.0%) and 'university undergraduate degree' (37.6%), indicating that the sample was quite educated. With respect to their primary role in the organization, 65.1% of the participants indicated that they were analysts, followed by manager of analysts (16.5%), other (15.6%), and analytic trainer/methodologist (2.8%). Among the analysts, 91.5% worked in an all-source environment and 5.6% worked in a single source environment. In terms of level of intelligence supported, 76.1% of the analysts indicated that they currently worked in a strategic environment, 12.7% worked in an operational environment, and 11.3% worked in a tactical environment.

4.2.2 Measures

The materials administered in the present study consisted of the following measures:

4.2.2.1 ICD 203 Scales

Two 13-item scales were developed by the second author to assess:

a) Attitudinal support for the facets of the ICD 203 analytic standard (ICD 203 Professional Values Scale [henceforth, ICD203-PVS]); and

b) Beliefs about organizational compliance with the same set of facets covered in the standard (ICD 203 Organizational Compliance Scale [henceforth, ICD203-OCS]).

For both scales, participants rated the extent to which they agreed with each of the statements on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree).

[2] The role of the managers and trainers was to communicate the information about the study to their staff; they did not have any direct involvement with scheduling participants or running participants, they were unaware of the names of the participants who did or did not participate, and they did not have access to the data.

The wording of the latter part of each of the items was identical across the two scales. In ICD203-PVS, the first part of the item was worded to reflect personal commitment to the standards. The specific items of ICD203-PVS are as follows:

1) It is important to me that I perform my analytic and informational functions from an unbiased, objective perspective.
2) It is important to me that I provide assessments that are not distorted by or altered with the intent of supporting a particular policy or political viewpoint.
3) It is important to me that I deliver analytic products in a timely manner, allowing them to be actionable by consumers.
4) It is important to me that I am informed by all relevant information that is available to the analytic element in my analyses.
5) It is important to me that I properly describe the quality and credibility of underlying sources in my analytic products.
6) It is important to me that I properly caveat and express uncertainties or confidence in analytic judgments in my analytic products.
7) It is important to me that I properly distinguish between underlying intelligence and my (or others') assumptions and judgments in my analytic products.
8) It is important to me that I incorporate alternative analysis where appropriate (e.g., explaining the strengths and weaknesses of alternative hypotheses in light of both available information and information gaps) in my analytic products.
9) It is important to me that I demonstrate relevance to Canadian national security in my analytic products.
10) It is important to me that I present analytic products in ways that facilitate clear understanding of the information and reasoning underlying my analytic judgments.
11) It is important to me that I deliver a key message that is either consistent with previous production on the topic or, if the key analytic message has changed, the product will highlight the change and explain its rationale and implications in my analytic products.
12) It is important to me that I make the most accurate judgments possible given the information available.
13) It is important to me that I deliver analytic products that effectively incorporate visual information where appropriate.

In ICD203-OCS, the first part of the item was worded to reflect the participant's view of their organization's commitment to the standards. However, the content of the 13 items was otherwise matched to ICD203-PVS. The specific items in ICD203-OCS are as follows:

1) In my organization, analysts and managers perform their analytic and informational functions from an unbiased, objective perspective.
2) In my organization, analysts and managers provide assessments that are not distorted by or altered with the intent of supporting a particular policy or political viewpoint.
3) In my organization, analytic products are delivered in a timely manner, allowing them to be actionable by consumers.
4) In my organization, analysis is informed by all relevant information that is available to the analytic element.

5) In my organization, analytic products properly describe the quality and credibility of underlying sources.

6) In my organization, analytic products properly caveat and express uncertainties or confidence in analytic judgments.

7) In my organization, analytic products properly distinguish between underlying intelligence and analysts' assumptions and judgments.

8) In my organization, analytic products incorporate alternative analysis where appropriate (e.g., explaining the strengths and weaknesses of alternative hypotheses in light of both available information and information gaps).

9) In my organization, analytic products demonstrate relevance to Canadian national security.

10) In my organization, analytic products are presented in ways that facilitate clear understanding of the information and reasoning underlying analytic judgments.

11) In my organization, analytic products deliver a key message that is either consistent with previous production on the topic or, if the key analytic message has changed, the product will highlight the change and explain its rationale and implications.

12) In my organization, the analytic element makes the most accurate judgments possible given the information available.

13) In my organization, the analytic products effectively incorporate visual information where appropriate.

4.2.2.2 Job Satisfaction

The following item taken from the 2015 US Federal Employee Survey (see Ref. [18]) was used to measure job satisfaction: "Considering everything, how satisfied are you with your job?" Responses to this item were made on a 5-point Likert scale ranging from 1 (*very dissatisfied*) to 5 (*very satisfied*).

4.2.2.3 The Big Five Inventory – Conscientiousness Subscale

The Big Five Inventory (BFI) [19] is a 44-item self-report inventory designed to measure the following five dimensions of personality: Openness, Conscientiousness, Agreeableness, Extraversion, and Neuroticism. Participants respond to each item using a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Items are written in both directions and reversal items are reverse scored. All subscales of the BFI are well known and are highly reliable; high Cronbach's alpha coefficients have been found even when the scale is administered to individuals from different cultures and/or in different languages [20], [21]. In this study, we collected data on the nine items of the Conscientiousness Subscale only. The Conscientiousness Subscale measures a tendency to be orderly, responsible, and dependable (e.g., "I see myself as someone who does a thorough job"). Responses to this item were made on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

4.2.2.4 The Actively Open-Minded Thinking Scale

The 8-item AOT scale [16] assesses a style of reasoning that includes the tendency to revise one's beliefs in response to new information and to take into consideration evidence that goes against their beliefs. The AOT has been found to correlate with various measures of reflective thinking [16]. Responses to these items were made on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

4.2.2.5 Organizational Commitment Scale

Organizational Commitment was measured using Allen and Meyer's [17] Affective Commitment Scale (ACS), Continuance Commitment Scale (CCS), and Normative Commitment Scale (NCS). The three components of organizational commitment are defined as follows:

a) Affective refers to employees' emotional attachment to, identification with, and involvement in, the organization (e.g., "I would be very happy to spend the rest of my career with this organization");

b) Continuance refers to commitment based on the costs that employees associate with leaving the organization (e.g., "It would be very hard for me to leave my organization right now, even if I wanted to"); and

c) Normative refers to employees' feelings of obligation to remain with the organization (e.g., "I think that people these days move from company to company too often" [17]).

The affective, continuance, and normative subscales each consist of eight items; the overall commitment scale consists of all 18 items. These items were rated on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

4.2.3 Procedure

The data was collected online anonymously using the Qualtrics survey platform during a single data collection session that lasted approximately 15 minutes. Participants received a link to the experiment through an email, which they accessed using a standard computer workstation.[3] To start the study, participants clicked on the link provided to them and read the background information. After providing consent, participants completed their ratings of ICD203-PVS and ICD203-OCS. Order of the two ICD 203 scales was counterbalanced across participants. Next, participants completed ratings on the job satisfaction item, the Conscientiousness subscale, the AOT scale, and the Organizational Commitment Scale. The order of these scales was randomized across participants. Lastly, participants provided demographic information.

4.3 RESULTS

4.3.1 Scale Characteristics

Psychometric analyses of the scales were conducted and reliabilities[4] were all acceptable (see Table 4-1), with the exception of the AOT scale which fell just below the criteria of acceptability, a finding that is not unusual for this scale [22].

**Table 4-1: Means, Standard Deviations and Reliability Coefficients for All Scales.
Note: *7-Item Scale: 1 = Strongly Disagree, 7 = Strongly Agree;
5-Item Scale: 1 = Strongly Disagree, 5 = Strongly Agree.

Scale	N	Mean	SD	Cronbach's Alpha
ICD203-PVS*	108	6.45	.49	.85
ICD203-OCS*	108	5.58	.83	.88
Job Satisfaction**	100	4.23	1.02	–

[3] While the majority of participants accessed this link using a workstation of their choice and during their own time, some CAF IC participants from within the Greater Toronto Area were brought into the lab at Defence Research and Development Canada (DRDC) – Toronto Research Centre (TRC) to participate in this study.

[4] In general, levels of Cronbach's alpha above 0.70 are considered to be acceptable and .80 are considered good (e.g., Refs. [23], [24]).

Scale	N	Mean	SD	Cronbach's Alpha
Conscientiousness**	106	4.26	.51	.77
AOT**	106	4.28	.49	.66
Affective Commitment*	101	4.57	1.00	.72
Continuance Commitment*	101	3.89	1.18	.79
Normative Commitment*	101	3.86	.96	.75
Overall Commitment*	101	4.11	.65	.74

Next, we conducted an exploratory factor analysis on ICD203-PVS responses. As Table 4-2 shows, we found a three-factor solution with eigenvalues of 5.10, 1.64, and 1.03, accounting for 39.26%, 12.64%, and 7.90% of the variance in ICD203-PVS, respectively. Together the three factors accounted for 59.80% of the variance in responses. A principal component analysis using Varimax rotation with Kaiser normalization converged in six iterations.

As Table 4-2 shows, the four items that loaded onto the first factor refer to the ability to remain unbiased (*unbiased*, *objective*; *not distorted/altered to support a view*; *express uncertainties or confidence in judgments*; *makes the most accurate judgments possible*). The five items that loaded onto the second factor represent items that refer to the rigor of the analysis (*describe source quality and credibility*; *distinguish intelligence vs. analysts' assumptions*; *incorporate alternative analysis where appropriate*; *present in ways to facilitate clear understanding*; *incorporates visual information where appropriate*). The four items that loaded onto the third factor refer to relevance of the analysis (*delivered in a timely manner*; *informed by all relevant information*; *demonstrate relevance to Canadian national security*; *deliver a key message that is consistent with previous or highlights changes*). Accordingly, we assigned the labels Unbiased, Rigorous, and Relevant, to Factors 1, 2, and 3, respectively.

Table 4-2: Factor Loadings of the ICD 203 Scale – Self-Perspective Condition.

Item Number and Wording		Factor		
		1 Unbiased	2 Rigorous	3 Relevant
1	Unbiased, objective perspective	0.84	0.17	0.13
2	Not distorted/altered to support a view	0.79	0.04	0.14
3	Delivered in a timely manner	0.4	0.15	0.64
4	Informed by all relevant information	0.43	-0.03	0.57
5	Describe source quality & credibility	0.19	0.84	0
6	Express uncertainties or confidence in judgments	0.59	0.21	0.46
7	Distinguish intelligence vs analysts' assumptions	0.42	0.61	0.08
8	Incorporate alternative analysis where appropriate	0.1	0.75	0.21
9	Demonstrate relevance to Canadian national security	0.06	0.31	0.71
10	Present in ways to facilitate clear understanding	0.27	0.48	0.41
11	Deliver a key message that is consistent with previous or highlights changes	0.13	0.15	0.64
12	Makes the most accurate judgments possible	0.78	0.18	0.28
13	Incorporates visual information where appropriate	-0.1	0.64	0.27

4.3.2 Responses to ICD 203 Scales

Figure 4-1 shows mean attitudinal endorsement of ICD203-PVS items with 95% confidence intervals. As can be seen, the overall level of endorsement approached the maximum scale value, indicating that participants strongly endorsed the standards. Using a Newman-Keuls test, mean scores for the four items with the highest means (accurate, timely, objective, not distorted to support political view) were significantly greater than mean scores for the four items with the lowest means (uses alternate analysis, incorporates visual information, describes quality and credibility, has relevance to Canadian national security). The mean score for the item uses alternate analysis was significantly lower than that of all of the other items except the item incorporates visual information. Other item means did not differ significantly from one another.

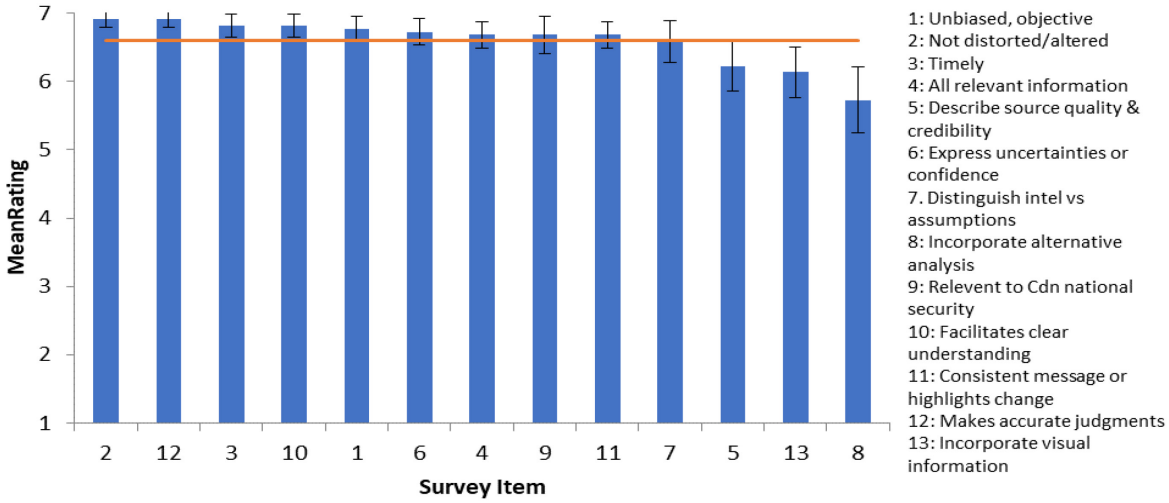


Figure 4-1: Mean Ratings on ICD203-PVS. Reference line shows grand mean.

As Figure 4-2 shows, on average, professionals agreed to some extent that all facets of ICD203 were addressed by their organization. A Newman-Keuls test revealed that the mean score for the item uses alternate analysis was significantly lower than that of all of the other items. Similarly, the mean score for the item describes source quality and credibility was also significantly lower than that found for each of the 11 items that had higher mean scores. The mean score for the item not distorted/altered to support a view was significantly higher than that found for each of the six items with the lowest mean scores (i.e., express uncertainties or confidence in judgments and items to the right of that in Figure 4-2), while the mean score for the item makes accurate judgments was significantly higher than that found for the four items with the lowest mean scores. Other item means did not differ significantly from one another.

To explore the extent to which there are perceived differences between analysts' internalized valuation of the ICD 203 analytic standards and their external evaluation of their organization's compliance with the standards, we calculated difference scores subtracting ICD203-OCS scores from ICD-PVS scores. As Figure 4-3 shows, all difference scores except for the item incorporates visual information significantly differed from 0, indicating that professionals endorsed the standards more strongly than they perceived their organizations as complying with them. Furthermore, results of the Newman-Keuls test revealed that the mean difference score for the item incorporates visual information was significantly lower than the mean difference score obtained for each of the other items. In addition, the three items with the largest disparity in mean scores had significantly greater difference scores than the four items with the smallest disparity in mean score. The mean difference score for the item uses alternative analysis, the item with the largest disparity, was also significantly greater than the mean difference score for the item makes accurate judgments. None of the other item means differed significantly.

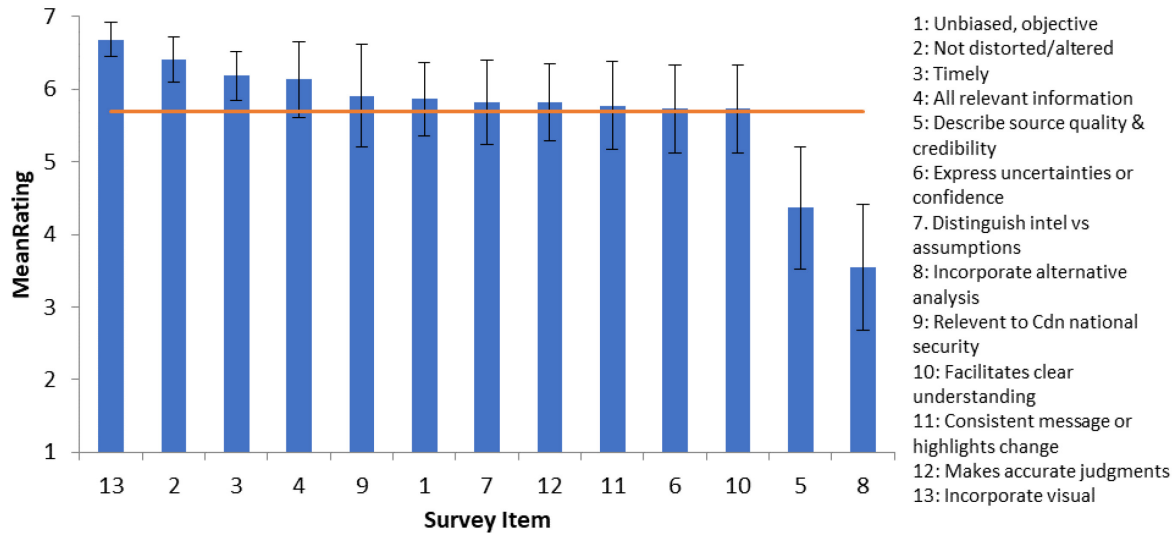


Figure 4-2: Mean Ratings on ICD203-OCS. Reference line shows grand mean.

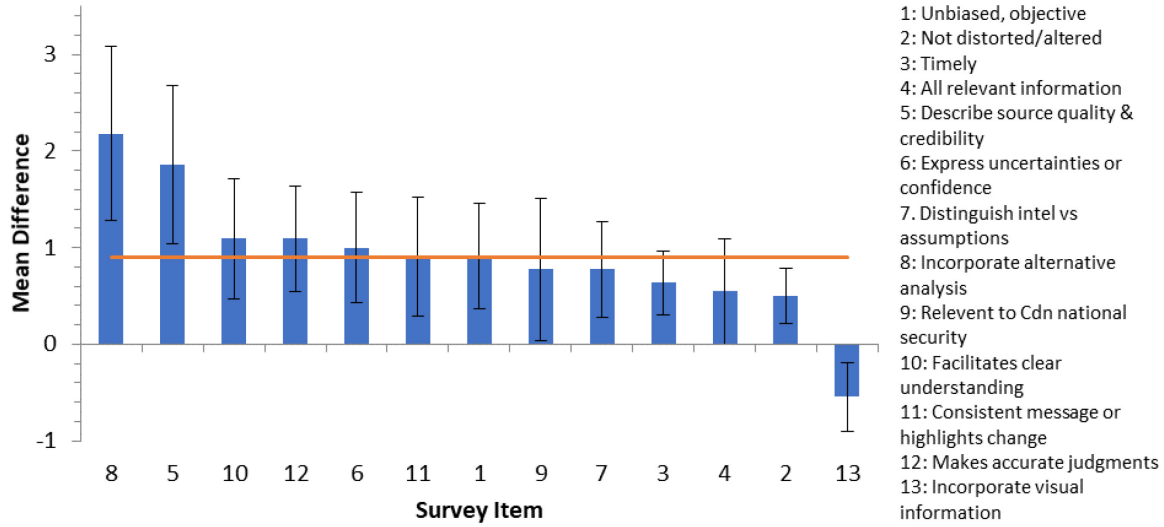


Figure 4-3: Mean Difference in ICD 203 Scores (ICD203-PVS – ICD203-OCS). Reference line shows grand mean.

4.3.3 ICD 203 Scale Correlates

Table 4-3 shows the correlations among the ICD 203 scales, includes the three ICD203-PVS factor scores, and the other measures taken in this study. As predicted, ICD203-PVS was significantly correlated with both conscientiousness and AOT. Conscientiousness was significantly correlated with each of the three factors and AOT was significantly correlated with the Unbiased and Rigorous factors. To test the hypothesis that analysts who score higher on conscientiousness and AOT will display stronger internalized values of analytic rigor, we conducted a regression analysis predicting ICD203-PVS scores from conscientiousness and AOT scores. Both the conscientiousness scale ($\beta = .45$, $t = 6.06$, $p < .001$) and the AOT scale ($\beta = .35$, $t = 4.61$, $p < .001$) significantly predicted ICD203-PVS, and the overall model was significant, $F(2, 102) = 37.24$, $R^2 = .42$, $p < .001$.

Table 4-3: Correlates of ICD 203 Scales. *p < 0.05, **p < 0.01.

	ICD203-OCS	ICD203-PVS	Unbiased	Rigorous	Relevant
ICD203-OCS	1				
ICD203-PVS	.40**	1			
Unbiased	.17	.51**	1		
Rigorous	.34**	.66**	.00	1	
Relevant	.17	.55**	.00	.00	1
Job Satisfaction	.39**	.11	.13	.12	-.05
AOT	.02	.46**	.40**	.29**	.10
Conscientiousness	.41**	.55**	.36**	.35**	.24*
Affective Commit.	.35**	.18	.06	.18	.06
Cont. Commit.	-.01	.05	-.13	.07	.13
Norm. Commit.	.23*	-.05	-.13	-.09	.14
Overall Commit.	.29**	.10	-0.11	.09	.17

We also examined the relation between ICD203-OCS and the individual difference measures taken in this study. As seen in Table 4-3, and confirming our hypotheses, ICD203-OCS scores were positively correlated with job satisfaction, affective commitment, and normative commitment, but did not correlate significantly with continuance commitment. ICD203-OCS was also positively correlated with AOT.

To further explore the relation between ICD203-OCS, job satisfaction, AOT, and commitment, we ran a regression analysis predicting ICD203-OCS scores from job satisfaction, conscientiousness, AOT, affective commitment and normative commitment scores. While job satisfaction ($\beta = .21$, $t = 2.06$, $p = .04$) significantly predicted ICD203-OCS, the other predictors did not account for a significant amount of variability in ICD203-OCS. The overall model was significant, $F(4, 95) = 5.62$, $R^2 = .44$, $p < .001$.

4.4 DISCUSSION

Intelligence organizations have long sought to promote analytic rigor. Over roughly the past decade, the US IC has formalized such efforts and issued directives to the IC's members on what facets of rigor are most important. Yet with few exceptions [11], [25], there has been little research examining how professionals respond to such directives, and the studies that have been conducted have been principally focused on the reliability of scoring the ICD 203 criteria. In the present research, we asked two, more basic questions: To what extent do intelligence professional endorse the contents of ICD 203 as valid professional values worth aspiring to and, secondly, to what extent do they view their organizations as compliant with such directives?

The results of the study revealed several important findings. First, we verified that the scales developed to address these questions had good psychometric properties. Both scales displayed very good reliability, which bodes well for future use. Moreover, the exploratory factor analysis of the attitudinal scale yielded a solution that not only explained a substantial proportion of the variability in responses to the items, but which reveals a coherent structure in which about one third of the items focus on being unbiased in one's thinking and judgments, another third of the items focus on being rigorous in one's thinking and judgments, and the remaining third focuses on striving for relevance to intelligence consumers. These higher-order concepts help to frame what ICD 203 is "about".

A second key set of findings reveal a very high degree of support, in general, for the ICD 203 facets, and although participants tended to view their organizations as complying with the facets of the standard, they judged compliance as falling somewhat short of the level of support for the facets that they personally expressed. While at face value the finding suggest that professionals are more willing to follow the directives of ICD 203 than their organizations manage to achieve, it may also represent a case of self-enhancement. That is, it is quite common for individuals to regard themselves as better than average on a range of self-relevant measures [26], and this may be yet another example of the sort of rose-coloured self-enhancing perceptions most people are prone to experience [27].

Third, although participants tended to support all of the ICD 203 facets, the differences in support were informative. All of the most highly endorsed items on ICD203-PVS refer to *aims* of intelligence such as being accurate, timely, and unbiased. In contrast, the items that were least strongly endorsed refer to *means* of improving intelligence analysis such as using alternative analysis, incorporating visual information into analysis, and describing source quality and information credibility. Evidently, intelligence professionals agree more strongly with the aims and objectives that motivate directives for promoting analytic integrity than they do with the specific means proposed for bringing about such objectives. These findings cohere with recent critiques of tradecraft methods that emphasize the scientifically undocumented nature of their effectiveness and their conceptual shortcomings (e.g., Refs. [28], [29], [30]), as well as with findings that analysts often do not use recommended methods such as structured analytic techniques because they remain unconvinced of their effectiveness [31] – a scepticism that recent empirical evidence suggests is indeed healthy and not unwarranted [32].

The present findings also offered support for our hypotheses regarding the correlates of the two ICD 203 scales. As we predicted, support for the facets of ICD 203 were positively correlated with conscientiousness and a disposition towards actively open-minded thinking. Although our study cannot establish causation, we believe it is more plausible that greater support for the facets of ICD 203 is attributable to dispositions such as conscientiousness and an actively open-minded thinking style than the other way around. Indeed, the pattern of correlation between these dispositions and the factor scores strengthens our conviction since the first two factors (Unbiased and Rigorous), which specifically pertain to facets of thinking style, are significantly correlated with AOT, whereas the third factor (Relevant), which does not pertain to thinking style, does not correlate with AOT. By comparison, we might expect conscientious individuals to score higher on each of the three facets, and indeed we observed significant correlations between conscientiousness and each of the three factor scores. Taken together, the correlations just noted have practical implications for how intelligence organizations could screen and select analysts. We already know from prior research that individuals who score higher in AOT tend to judge more accurately (e.g., Refs. [15], [33]). The present findings indicate that people who score high in AOT also aspire to be more accurate, unbiased, and rigorous in their thinking and judgments. Therefore, intelligence organizations might consider screening analysts for dispositions such as conscientiousness and AOT, giving more weight in selection processes to those who score highly on these measures.

The causal direction of our second set of correlates is harder to decipher. We found that professionals who were more satisfied, affectively and normatively committed and higher in AOT judged their organization to be in stronger compliance with the directives in ICD 203. Our multiple regression analysis predicting ICD203-OCS from those significant correlates yielded a single significant predictor – job satisfaction. One plausible explanation is that professionals who are relatively more satisfied with their jobs view their organization more favourably as a result of attribute substitution [34]. In effect, when asked how good their organization is performing in various respects, participants might use how good they feel about their job as a cue to answering the question. However, it is also plausible that ratings of job satisfaction factor in to some extent how well professionals judge their organizations to be meeting the high-level goals of their profession. Future research might attempt to capitalize on organizational interventions aimed at bolstering analytic integrity to gauge whether such fortuitous "manipulations" simultaneously trigger improved job satisfaction, organizational commitment, as well as concomitant improvements in real and perceived analytic integrity at an organizational level.

4.5 CONCLUSION

The present research developed two scales for measuring attitudinal support for, and perceived organizational compliance with, the facets of ICD 203, the US IC's directive on analytic integrity. The scales developed were found to be reliable and coherently structured. Intelligence professionals in Canada, who are not formally bound by the ICD 203 directives, and who may not even be aware of their existence, nevertheless strongly supported all facets of the directive. They were particularly supportive of aspects that refer to aims or objectives of intelligence production, and they were correspondingly less supportive of specific means for achieving those goals. However, in absolute terms, support was strong across the full set of items. Professionals also regarded their organizations as complying well with the facets of ICD 203. Beyond shedding light on the responses of this sample of intelligence professionals, the research has spurred an interest by at least one Canadian intelligence organization to proactively use the ICD 203 scales to monitor its members' attitudes towards analytic integrity and their perceptions of the organization's compliance with such standards on an annual or semi-annual basis. As such, the ICD 203 scales may serve as a useful, low-cost source of information for promoting situational awareness and accountability in intelligence organizations.

4.6 REFERENCES

[1] Borek, J.J. (2017). Analytic tradecraft in the U.S. intelligence community. Doctoral dissertation. Minneapolis, MN: Walden University. https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=5081&context=dissertations.

[2] Marchio, J. (2014). Analytic tradecraft and the intelligence community: Enduring value, intermittent emphasis. *Intelligence and National Security* 29(2):159-183.

[3] Tobin, Y. (2005). The reorganization of the intelligence community. *United States Attorneys' Bulletin* 53(4):2-7.

[4] PUBLIC LAW 108-458 108th Congress. (2004). Intelligence Reform and Terrorism Prevention Act. Retrieved from http://www.gpo.gov/fdsys/pkg/PLAW-108publ458/pdf/PLAW-108publ458.pdf.

[5] Office of the Director of National Intelligence. (2007). Intelligence community directive number 203: Analytic Standards. Retrieved from https://www.dni.gov/files/documents/ICD/ICD%20203%20Analytic%20Standards%20pdf-unclassified.pdf.

[6] Pigg, V.H. (2009). Common analytic standards: Intelligence community directive #203 and US Marine Corps intelligence. *Small Wars Journal*. Retrieved from http://smallwarsjournal.com/blog/journal/docs-temp/260-pigg.pdf.

[7] Office of the Director of National Intelligence. (2015). Intelligence Community Directive Number 203: Analytic Standards. Retrieved from https://www.dni.gov/files/documents/ICD/ICD%20203%20Analytic%20Standards.pdf.

[8] Rojas, Lt. Col. J.T. (2016). Masters of analytical tradecraft: Certifying the standards and analytic rigor of intelligence products. Maxwell Air Force Base Research Report. Montgomery, AL: Maxwell Air Force Base.

[9] McGlynn, P., and Garner, G. (2018). *Intelligence Analysis Fundamentals*. Boca Raton, FL: CRC Press.

[10] Cardillo, R. (2010). Intelligence community reform – A cultural evolution. *Studies in Intelligence* 54 (3):1-7 (Extracts, September 2010).

[11] Marcoci, A., Burgman, M., Kruger, A., Silver, E., McBride, M., Singleton Thorn, F., Fraser, H., Wintle, B.C., Fidler, F., and Vercammen, A. (2019). Better together: Reliable application of the post-9/11 and post-Iraq US intelligence tradecraft standards requires collective analysis. *Frontiers in Psychology* 9:2634.

[12] Roberts, B.W., Jackson, J.J., Fayard, J.V., Edmonds, G., and Meints, J. (2009). Conscientiousness. In: *Handbook of Individual Differences in Social Behaviour,* Leary, M.R., and Hoyle, R.H. (Eds.), 369-381. New York, NY: Guilford Press.

[13] Baron, J. (1985). What ends of intelligence components are fundamental? In: *Thinking and Learning Skills*, Chipman, S.F., and Segal, J.W. (Eds.), 2:365-390. Hillsdale, NJ: Lawrence Erlbaum.

[14] Baron, J. (1993). Why teach thinking? An essay. *Applied Psychology.* 42(3):191-214.

[15] Haran, U., Ritov, I., and Mellers, B.A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making* 8(3):188-201.

[16] Baron, J., Scott, S., Fincher, K., and Metz, S.E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition,* 4(3):265-284.

[17] Allen, N.J., and Meyer, J.P. (1990). The measurement and antecedents of affective, continuance and normative commitment to the organization. *Journal of Occupational Psychology,* 63(1):1-18.

[18] Office of Personnel Management. 2015 United States Federal Employee Survey. Retrieved from https://www.fedview.opm.gov/2015/What/.

[19] John, O.P., and Srivastava, S. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives. In: *Handbook of Personality: Theory and Research*, Pervin, L.A., and John, O.P. (Eds.), 2:102-138. New York, NY: Guilford Press.

[20] McCrae, R.R., Costa, P.T., Del Pilar, G.H., Rolland, J., and Parker, W.D. (1998). Cross-cultural assessment of the five-factor model: The revised NEO personality inventory. *Journal of Cross-Cultural Psychology,* 29 (1):171-188.

[21] Ortiz, F.A., Church, A.T., Vargas-Flores, J.D.J., Ibáñez-Reyes, J., Flores-Galaz, M., Iuit-Briceño, J.I., and Escamilla, J.M. (2007). Are indigenous personality dimensions culture-specific? Mexican inventories and the five-factor model. *Journal of Research in Personality,* 41(3):618-649.

[22] Mandel, D.R., and Kapler, I.V. (2018). Cognitive style and frame susceptibility in decision-making. *Frontiers in Psychology,* 9:1461.

[23] Loewenthal, K.M. (2004). *An Introduction to Psychological Tests and Scales*, 2nd ed. Hove, UK: Psychology Press.

[24] Nunnally, J.C. and Bernstein, I.R. (1994). *Psychometric Theory*. New York, NY: McGraw-Hill.

[25] Marcoci, A., Vercammen, A., and Burgman, M. (2019). ODNI as an analytic ombudsman: Is Intelligence community directive 203 up to the task? *Intelligence and National Security,* 34(2):205-224.

[26] Alicke, M.D., and Govorun, O. (2005). The better-than-average effect. In: *The Self in Social Judgment*, Studies in Self and Identity Series, Alicke, M.D., Dunning, D.A., and Krueger, J.I. (Eds.), 85-106. New York, NY: Psychology Press.

[27] Taylor, S.E. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin,* 110(1):67-85.

[28] Chang, W., Berdini, E., Mandel, D.R., and Tetlock, P.E. (2018). Restructuring structured analytic techniques in intelligence. *Intelligence and National Security,* 33 (3):337-356.

[29] Dhami, M.K., Mandel, D.R., Mellers, B.A., and Tetlock, P.E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science,* 10(6):753-757.

[30] Mandel, D.R., and Tetlock, P.E. (2018). Correcting judgment correctives in national security intelligence. *Frontiers in Psychology*, 9:2640.

[31] Coulthart, S. (2016). Why do analysts use structured analytic techniques? An in-depth study of an American intelligence agency. *Intelligence and National Security,* 31(7):933-948.

[32] Mandel, D.R., Karvetski, C. and Dhami, M.K. (2018). Boosting intelligence analysts' judgment accuracy: What works, what fails? *Judgment and Decision Making,* 13(6):607-621.

[33] Tetlock, P.E. (2005). Expert political judgment: *How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.

[34] Kahneman, D., and Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In: *Heuristics and Biases: The Psychology of Intuitive Judgment*, Gilovich, T., Griffin, D., and Kahneman, D. (Eds.), 49-81. New York, NY: Cambridge University Press.

Chapter 5 – INTRODUCING AN EVIDENCE-BASED APPROACH TO ANALYTICAL TRADECRAFT TRAINING

Kathryn, Kate, Jo, Sam and Andy
UK Analytical Tradecraft Training Team
UNITED KINGDOM

5.1 INTRODUCTION

In 2003, the UK was at war in Iraq. This war was reportedly based, in part, on intelligence which suggested Iraq was in possession of Weapons of Mass Destruction. No such weapons were found and the subsequent inquiry into this event [1] concluded that the key intelligence used to justify the war was much less reliable than it had been interpreted. One of the report's recommendations was that the intelligence community become more rigorous in its analytical methods so that intelligence was less likely to be overly influenced by what customers wanted to hear. To the UK intelligence community this report was a catalyst to reform many aspects of intelligence reporting. This is the story of how one of the UK's Intelligence Agencies transformed its approach to analytical thinking by reaching out to academia and commissioning their own research, and starting to embed an evidence-based approach to tradecraft in their organisation.

5.2 A BRIEF HISTORY OF OUR ANALYTICAL THINKING TRAINING

In the 13 years between 2005 and 2018, one of the UK's largest intelligence agencies has been through six significant step changes in how analytical thinking is taught, assessed and recognised. In 2005 the predominant handbook for intelligence analysts was Heuer's *Psychology of Intelligence Analysis* [2]. It was from this work of former CIA analyst Richards Heuer that Generation 1 analytic thinking tradecraft training was born. A group of experienced analysts working on Counter-Terrorism designed a workshop affectionately known as 'Thinky' to communicate the key themes of Heuer's work to analysts across the wider Counter-Terrorism team. Thinky asked analysts to think about their thinking by spending some time introspecting and recognising their cognitive biases. Thinky was later adapted and extended to the wider organisation and partner agencies under Generation 2 – Thinky for Everyone.

It was down to the success and reach of Thinky that analysts who recognised their bias had started searching for solutions. This is where Generation 3 – The Four Pillars Learning Path emerged. Another group of experienced analysts had developed a model to describe analytical tradecraft they called 'the four pillars of analysis'. The four pillars were psychology, creativity, structure and challenge. Four groups of experienced analysts took one pillar each and developed a series of courses designed to deep-dive into each of these aspects. This phase was heavily influenced by the Sherman Kent School of Intelligence Analysis perspective on Structured Analytic Techniques (SATs), and the format was based on their in-house training programmes.

After a couple of years of delivering The Four Pillar Learning Path, it had become clear that in an effort to recognise four complementary aspects of analysis separately, new analysts going through the curriculum were missing the power of combining the separate pillars. Generation 4 – Analytical Thinking Parts 1 and 2 combined the separate classes into a much more coherent course.

In 2014 – 2015, the Agency was planning to recruit and train an influx of new analysts and our analytical thinking training needed to be updated to serve the needs to two distinct customers, brand new analysts, and experienced analysts. There was also a drive to train new analysts at Unclassified level. Separately, some members of the training team had been working with external researchers to better understand the science of intelligence analysis, and take their theoretical knowledge beyond Heuer's 1999 book. These drivers

converged with the launch of Generation 5 – Analytical Tradecraft Foundations and Advanced Analytical Tradecraft courses. These two courses reduced the focus on SATs and increased focus on underlying good practice principles from academia and from across the intelligence community.

The more evidence-based, principles-focused curriculum was working well for new analysts. However, we consistently found that by stripping out the classified intelligence context for the Foundations course, analysts were finding it difficult to translate the generic principles of the analytic workflow to their work. Generation 6 – Analytical Thinking Workshop was developed to bring the science behind the analytical workflow together with recent operational application. Generation 6 is already proving to be hugely effective, having built on the previous five generations on training that came before. Table 5-1 provides more detail on each of these iterations giving a flavour of what the training curriculum looked like at each of these points.

Table 5-1: The Evolution of Analytic Tradecraft Training.

Generation	Dates	Programme Outline
Generation 1: Thinky for Counter-Terrorism (CT)	2006 – 2008	A 2-day course. The first day consisting of a practical analytic scenario, the second day de-brief and introduction of psychology of intelligence analysis topics. The aim was to introduce analysts to the theory of cognitive biases and get them to reflect on how their approach their own work may be influenced by thinking heuristics.
Generation 2: Thinky for Everyone	2008 – 2009	A 2-day course. This was essentially the same course as Thinky Generation 1, but the team intentionally widened the pool of adjunct trainers beyond the CT Team and encouraged trainers to bring stories and examples of how they had seen cognitive biases at work.
Generation 3: The Four Pillars Learning Path	2009 – 2012	A 10-day-long learning path analysts could take over a recommended 6-month period. The first course was Thinky re-branded as a 2-day Psychology class. This was followed (in any order) by a 2-day Structure class, 1-day Creativity class, 1-day Challenge class, and a 2-day SATs Facilitation class. The learning path was concluded by a 2-day Four Pillar Challenge course where analysts were given a day-and-a-half-long scenario to tackle in teams.
Generation 4: Analytical Tradecraft Parts 1 and 2	2012 – 2015	Part 1 was a 5-day class which combined elements from the previous learning path. It covered theory, tools and structured techniques from all four pillars of analytical thinking; psychology, creativity, structure and challenge. It was followed by Part 2 several months later which was the same as the Four Pillar Challenge course from Generation 3.
Generation 5: Analytical Tradecraft Foundations and Advanced Analytical Tradecraft	2015 – 2017	Two 5-day courses. The 5-day Unclassified Analytic Tradecraft Foundations course covered the principles of the 6-stage generic analytic workflow and some basic tools and tips for each stage. A classified 5-day Advanced Analytic Tradecraft course introduced more advanced SATs and was led by experienced analysts.

Generation	Dates	Programme Outline
Generation 6: Analytical Tradecraft Workshop	2017 – current	An e-learning module and a 5-day Bootcamp. An hour-long e-learning module introduced the generic analytic workflow and explained the research that led to its development. The 5-day Bootcamp covered a toolkit of tips, techniques and SATs that experienced analysts have found useful in applying the principles of the workflow.

5.3 AN EVIDENCE-BASED APPROACH TO INTELLIGENCE ANALYSIS

In our brief history above, we explained that the introduction of the science of intelligence analysis was the key influence in moving from Generation 4 to Generation 5 analytic tradecraft training. This section will cover how we applied an evidence-based approach to intelligence analysis via development of training and support aids. We will describe some of the original research that we commissioned and how this was exploited to create evidence-based analytic tradecraft training and guidance.

From the inception of analytical thinking training, the single biggest influence had been the 1999 book *The Psychology of Intelligence Analysis*, written by ex-CIA manager Richards J. Heuer. Heuer directly applies early work on cognitive biases to intelligence analysis and explains several biases and how they might present in intelligence analysis, before presenting a new SAT, the Analysis of Competing Hypotheses (ACH), as a method for mitigating (confirmation) bias and producing better intelligence. The ACH method has been picked up and applied across the intelligence community. However, we became acutely aware that there was little to no scientific basis for Heuer's claims that cognitive bias is rampant in intelligence analysis [3], or for his conclusion that ACH can improve analytic performance [4], [5]. This is something that Heuer himself recognised, resulting in the recommendation in the concluding chapters of his book that research should be done into these areas. Regrettably, very little academic research was conducted on the topic, leading the intelligence community to develop a negative view of analysts as being biased, and to attempts to cure them by telling them to use ACH.

In 2011, members of the Analytical Thinking Training Team began to search for more up-to-date academic research on the science of intelligence analysis to inform training programmes. We discovered that the vast majority of academic work available at the time on intelligence analysis was not scientifically rigorous. For example, it was based on case studies, on qualitative analyses of very small samples of analysts, or on larger samples of non-analysts. Often, the past work did not focus on representative analytic problems, and so it was unsurprising that it rarely produced learning that ended up being applied in the intelligence community. The research we needed was not out there and we concluded that we would have to go about recruiting expert help and conducting it ourselves.

A useful scientific perspective is research in the area of Judgement and Decision Making (JDM) (also called Decision Science). This is an inter-disciplinary science predominantly influenced by cognitive psychology and economics. JDM has been successfully applied to numerous areas such as the legal system, healthcare and public policy but had not previously been applied to intelligence analysis. With confidence that the science of JDM could be applied to intelligence analysis, we developed a two-pronged approach, conducting both bottom-up scientific research testing assumptions that underpin current policy and practice, as well as top-down empirical research on analytic tasks and analysts' performance. In contrast to the previous research on intelligence analysis, we decided to take a primarily quantitative approach, and made the most of our unique access to real intelligence analysts and the sorts of tasks they are presented with.

We formed a small virtual team to explore what we termed 'Analytic Decision Science' (ADS), borrowing from the fields of JDM, cognitive science and computer science to develop an evidence-based approach to policy and practice in intelligence analysis. The team focused on three areas: conducting original applied

research into the science of intelligence analysis, working with partners across the agency to exploit the findings from our research, and building an enduring ADS capability that could continue to provide benefit to the wider intelligence community by developing evidence-based good practice for solving problems and de-mystifying the craft of intelligence analysis.

Since 2011, working in partnership with a leading academic in Decision Science, Professor Mandeep K. Dhami, we have conducted several original research studies in the following key areas:

1) Analytic workflow and strategies [6], [7], [8];
2) SATs [4], [9], [5], [10];
3) Communication of uncertainty in intelligence analysis [11], [12];
4) Collaborative analysis [13]; and
5) Analyst's technical tools [14].

We have also conducted critical reviews of past research and existing policies and practices in the intelligence community pertaining to cognitive biases, SATs and communicating uncertainty [3], [15]. In this chapter, we primarily focus on our research on the analytic workflow and strategies. This includes identifying the generic analytic workflow and developing the Analysis Support Guide (ASG).

5.3.1 How Should, and How Do, Analysts Structure Their Work?

There have been several attempts at defining a workflow for intelligence analysis and we started out conducting a large-scale review of everything we could find on 'how' analysts should be structuring their analytic work. We reviewed prominent workflow models from academia, namely the 7-phase model of analysis [16], the three primary analytic cognitive functions approach [17], the 16-step model of analysis [18], the data/frame sense-making model [19], the 5-activity analytic model [20], and the 4-phase model of analysis [21]. We also reviewed a variety of classified and unpublished sources from across the intelligence community such as organisational skills and analyst development standards, tradecraft bulletins and working guides [22], [23].

We found that although the sources varied in the detail, two key themes emerged as common across all of them: good analysis should be *ordered* and it should involve some form of *critical thinking*. We also found that there were inconsistencies over the level of granularity. There was also a lot of alignment on the generic components of the workflow. Our goals differed from previous work where workflows had been constructed based on analyst technical tool use with a view to informing analytic tool development. Rather, we wanted to understand intelligence analysis from an analyst-centric point of view, at a level of granularity that could be observed, trained and assessed, and tool-agnostic. With this goal in mind, we constructed the Generic Analytic Workflow, or Generic Analytic Cycle [6]. It is generic enough to apply to all analytic problems and can be scaled both up and down. For more complex problems it will be highly iterative, and for simpler problems you may only spend seconds at each stage. The workflow consists of six stages, and these reflect necessary components on which later stages depend (i.e., you cannot plan effectively if you do not know what your intelligence requirements are). If you plan effectively, on the other hand, you will obtain relevant data, which will need to be processed, and then interpreted, and communicated – in that order. The six stages of the workflow are therefore:

1) Understand requirements;
2) Plan analytic response;
3) Obtain data;
4) Process data;
5) Interpret data; and
6) Communicate conclusions.

Table 5-2 provides more detail on what each stage entails.

Table 5-2: The Generic Analytic Workflow [6].

Stage 1 – **Understanding Requirements**. This stage is about understanding the customer's point of view, the wider context for the immediate requirement or intelligence question, and what the ultimate aim or outcome is and how it will be achieved.
Stage 2 – **Planning Your Analytic Response**. This stage is about identifying alternative methods that could be employed to fulfil the requirement, evaluating them in terms of how efficient and effective they may be and then making a prioritised plan for how to proceed.
Stage 3 – **Obtaining Data**. This stage is about extracting, filtering and selecting the relevant data from the most appropriate sources. It will involve being able to query multiple data sources in the most surgical and efficient way or establishing new sources of data if nothing currently exists.
Stage 4 – **Processing Data**. This stage is about understanding the raw data output of a tool, and being able to describe accurately (in plain English) what the data means. It may involve exporting data from multiple sources and reformatting it into composite charts, diagrams or other visualisations.
Stage 5 – **Interpreting Data**. This stage is about testing alternative explanations for the (often incomplete) 'facts', and constructing strong logical arguments to support conclusions as well as dismiss alternative hypotheses. It may involve distinguishing between different strands of activity.
Stage 6 – **Communicating Conclusions**. This stage is about presenting and communicating the outcome of analysis in a clear, meaningful, and relevant way. It will involve determining the appropriate medium to share conclusions, as well as highlighting and explaining areas of uncertainty.

By fusing intelligence community doctrine with academic research, we had constructed a single simple model to explain how analysis should be ordered. We then developed a series of observable behavioural indicators for each of the six stages which described how critical thinking applied at this stage. This became our 'normative' model for 'good' intelligence analysis, meaning our next logical question was a descriptive one.

To determine how analysts actually apply structure to their work, we designed a study to measure the extent to which analysts took time to understand the requirement before thinking about how to answer it; and how often they did that before collecting some data. We wanted to know whether they fully understood their data before 'connecting the dots', drawing conclusions and constructing arguments. We surveyed 144 analysts across a two-week period, with good representation across teams, experience, and skill levels [6]. The survey gave analysts a hypothetical scenario representative of a typical analytic task they would be asked to complete. This followed with six activities, each activity representing each stage of the workflow. Analysts ranked the activities in the order they would complete them in response to the given scenario. We measured how consistently analysts ordered the activities compared to the Generic Analytic Workflow.

We found that only 16% of analysts ordered the activities in the logical order suggested by the workflow. Another 7% ordered the first three phases in the desired way, and a further 7% ordered the last three phases appropriately. However, the vast majority of analysts surveyed (70.1%) applied less ordinal structure to their work. Of these analysts, we found that they tended to delay planning and/or start interpretation prematurely. We also found that application of ordinal structure was unrelated to experience, training received or corporately assessed skill level [6].

5.3.2 What Strategies Should, and What Strategies Do, Analysts Use to Solve Problems?

Next, we tackled the second key theme from our literature review, namely the extent to which analysts applied critical thinking to their work. Cognitive science has shown that critical thinking requires the use of System 2 or deliberative cognition [24], [25], [26], as opposed to System 1 or intuitive thinking. This is not to say that System 1 (intuitive thinking) does not have any value. Indeed, there are disagreements in academia over where and when intuitive thinking is appropriate. If critical thinking is required for good-quality intelligence analysis, then we should like to see evidence of analysts applying System 2 (deliberative thinking) when tackling analytic tasks.

It has been difficult to obtain a clear picture of how experienced analysts work with many previous studies using a think-aloud protocol or observational methods with small numbers of analysts [27], [28], [29], [30], [31]. None had specifically examined the application of deliberative and intuitive thinking at each stage of the workflow.

To investigate this issue, we surveyed 113 analysts with a range of experience and skill level [7]. The survey presented analysts with six hypothetical (but representative) scenarios signifying each stage of the generic analytic workflow, followed by a selection of activities. Analysts were asked to indicate how likely they would be to conduct each activity in response to the given scenario using a five-point scale labelled “never”, “a little”, “some”, “a lot”, and “always”. In each scenario, some of the activities represented deliberative thinking at this stage of the workflow, while others represented intuitive thinking. We measured the extent to which analysts said they would do deliberative or intuitive thinking activities at each stage. Due to the clear emphasis in intelligence community guidance for critical thinking in analysis we predicted that years of experience, skill level and analytical thinking training would be positively associated with application of deliberate thinking strategies.

We found that analysts said they would use deliberative strategies more often than intuitive ones at the initial and final stages of the workflow. Analysts also said they would use deliberative strategies more often than intuitive ones at the processing data stage. There was no significant difference in how often analysts said they would use intuitive and deliberative strategies at the plan analytic response, obtain data and interpret outputs stages of the workflow. Years of experience working in the intelligence community, skill level, analytic thinking training, and time spent working collaboratively (opposed to individually) were largely unrelated to reported strategy use [7].

5.3.3 Development and Validation of the Analysis Support Guide

Having discovered that application of ordinal structure and of critical thinking was not significantly related to analysts’ years of experience, training received or corporately assessed skill level [6], [7], our next goal was to develop an intervention could demonstrably increase analysts application of these two key components of good-quality intelligence analysis. In professional domains where judgements are typically based on partial and conflicting information, decision support guides have been shown to promote both intuitive and critical thinking, as well as consistency, transparency and accountability [32]. We therefore developed an Analysis Support Guide [8].

The ASG aims to capture, communicate and encourage good practice. It contains the generic analytic workflow, prompts for good practice at each stage of the workflow, indicators of good and poor analytic practice, and an analytic investigation questionnaire.

We conducted a small-scale content validation study where we briefed and surveyed fourteen intelligence professionals, a mix of analysts and managers of analysts [8]. We learned that the generic analytic workflow was something that was recognised as good practice by both analysts and managers. We had universal agreement on the criticality and relevance of the six distinct stages, and of the vast majority of the content of

the guide. There were several points where the wording led to ambiguity and this was amended for the final version. However, many of the participants in this study raised concerns about the paper-based ASG not being very user-friendly in its current paper format, and in recognition of this we intend to seek ways to seamlessly incorporate prompts, or 'nudges', into the analyst's immediate environment, for example, through the technical tools they use.

5.3.4 Testing the Analysis Support Guide

Having established the content validity of the ASG, we next wanted to find out if using it would increase the application of good practice, and thus increase the quality of analytic output. Therefore, we conducted a randomised controlled trial where we randomly allocated analysts to two groups and asked them to complete the same representative intelligence analysis task [33]. The experimental group of analysts were first given a brief on the ASG and were asked to use it to structure their work, while the 'active' control group of analysts received a brief on a different topic. Analysts worked alone and were asked to produce a written report communicating their findings as well as supporting documentation to 'show their workings'. Analysts then completed a survey which asked them to self-report how they had approached the task. The final reports were blind-marked for:

a) Indicators that ordinal structure and critical thinking as described in the ASG were applied;

b) Accuracy of the conclusions (which were available due to the specifically designed scenario); and

c) The level of 'customer impact' achieved, defined as having answered the customers most important needs in the required timescales.

Disappointingly, those asked to use the ASG did not always do so. We believe this could be attributable to the previously identified issue of the ASG in its paper form not being user-friendly [8], and this was corroborated by survey responses from participants in the experimental group, many of whom said they tried but found it difficult to apply the guidance in the ASG because there was too much information in it to digest and apply in too short a time. They said that not having had the time to assimilate the new information ultimately slowed them down and made them feel pressured. We learned that a 30-minute briefing on the ASG followed by a two-hour exercise was insufficient to test the ASG.

Nevertheless, an analysis of the data from across the two groups did yield some encouraging results. We simply ranked all of the participants by how well they applied good practice as identified in the ASG, and then compared the accuracy and impact scores for the top and bottom thirds. We found that the group who applied good practice produced significantly more accurate and significantly more impactful reports. However, for the reasons outlined above, and because it is good scientific practice to do so, our test of the ASG needs to be replicated.

5.3.5 Embedding Good Practice in Training and Assessment

In 2015, around the time we were concluding the research detailed above, our organisation was designing a structured development programme for new analysts. This presented an opportunity to build our latest research into our core organisational training and take a more problem-centric approach to analysis. The ADS research team worked with senior analysts and the analytical thinking training team to incorporate findings from the research outlined in this chapter into the next generation of analytical thinking training for analysts. Generation 5: Analytical Tradecraft Foundations and Advanced Analytical Tradecraft was created consisting of two 5-day courses both structured around the 6-stage analytic workflow. Some of the cognitive biases and SATs which had been taught in generations 1 – 4 were included in the new curriculum alongside up-to-date caveats drawn from Belton and Dhami's [3] review of cognitive biases in the intelligence analysis domain, and Dhami, Belton and Careless's [15] review of SATs.

Spurred by previous findings which showed that the level of analytical training analysts had received was not associated with the extent to which they applied an ordinal structure to their work or the extent to which used critical thinking in practice [6], [7], in mid-2016 the ADS team investigated how successful the new training programme had been at embedding good practices into intelligence analysts through training. We compared scores of analysts who had been through the new training and existing analysts who had not on their application of good practice, accuracy of the conclusions drawn and customer impact achieved. We found that on average analysts who had completed the training consistently outperformed existing analysts who had not. We would like to continue assessing and evaluating the efficacy of the evidence-based training programme.

5.4 CONCLUSION

Since we started on our journey to institute a more evidence-based approach to intelligence analysis, the need for analytical thinking training and recognition has continued to grow. Over recent years changes to the environment in which we work, and the challenges this poses, has made it harder for intelligence analysts of all kinds to do their critically important job. The changing environment also makes understanding the psychology of intelligence analysis even more important, as well as finding ways of helping analysts harness their creativity and creating time for analysts to apply an ordinal structure and critical thinking.

From 13 years of continuously iterating and improving our analytical thinking training programme and seven years of original research, we have demonstrated how intelligence analysis can be an evidence-based profession, in its' approach to skill and competency development as well as in the rigorous quality expectations of its output. We have demonstrated that intelligence analysis can be thought of as a skill that can be taught, assessed and improved. We have learned that good intelligence analysis is accurate, systematically ordered and critical, and has impact. We can think of these three aspects as: did we come to the right conclusions? Did we get there in a thoughtful systematic (ordered and critical) way? And did we deliver the intelligence in time and in a format that helps our customers achieve their goals? Understanding and applying the principles behind the six stages of the analytic workflow gives the best chance of creating great actionable intelligence that meets its customers' needs. Now that we know this, we have started to embed it into our enduring systems and processes. Analytical thinking and advanced analytical tradecraft are now taught in our foundation course for new analysts, and are built into our analysis skills framework. Perhaps most importantly we have ignited a desire within the analytical thinking training community for an evidence-based approach to tradecraft which we hope will continue to flourish.

Like any cultural shift, we faced a lot of challenges along our journey to evidence-based analytic tradecraft. In the early days finding what evidence base existed for intelligence analysis was a challenge due to the dearth of information available and limited applicability of its findings. When we started sponsoring our own research programme, gaining access to real analysts became the main challenge. They often have other priorities and gatekeeper managers are not interested in spending time on scientific inquiry. Once we had resolved those issues communicating the findings of our research in a way that tradecraft practitioners could, and would use became our next challenge. Often our research found things people did not want to hear and so the incentive to do what the science was showing was usually low. In the end, we found passionate supporters who understood both the science and the application and were able to do the translation. This, however, is not yet systemic, and we have been unable to secure a permanent foothold to continue sponsorship of analytical tradecraft research and pull-through which will survive key individuals moving postings.

Another set of challenges we had to deal with along the way related to the fact that intelligence organisations do not typically have the infrastructure required to conduct research that can be published in academic outlets. It was important for us to publish our work because we believed our work could be improved through the peer review process and because we believed our work could have wider impact through publication. Therefore, we had to develop a capability to allow us to do this. First, we commissioned the

establishment of an 'annotated' resource library on the psychology of intelligence analysis that could be used to support research on intelligence analysis. There are over 200 documents (and their summaries) in this library. Second, we developed an internal review process so that any research that was conducted (classified or not) could be released for publication. This process uses resources and is time-consuming but is necessary. The third aspect of the infrastructure, namely, the ethics of conducting research was initially overcome by having our academic collaborator's (namely, Professor Dhami's) institutional research ethics review board assess the research. We recognize it would be beneficial for intelligence organisations to have their own review boards and the intelligence community in the UK has been looking into various options.

Of course, other challenges also remain. Perhaps the most obvious one is that we need more research evidence! We are still teaching (an albeit smaller number) of SATs, selected based on the best evidence we have, but we need a better evidence base showing under what circumstances the numerous SATS are most effective, and what circumstances they do not help. This research is beginning to emerge [4], [5]. We need academics that understand the world of intelligence analysis. Fortunately, decision scientists in the UK, Canada and the US are heralding a new era of research on intelligence analysis and in doing so are holding the intelligence community to account [34]. Relatedly, in order to bring about change in policy and practice, we also need intelligence agencies that are able to understand the latest scientific advances, and are committed to pulling through insights and findings on a continual basis. Again, there are signs that this is beginning to emerge. The ADS team (we) are one example. We have not only impacted our own organisation, but our work has also influenced parts of the worldwide intelligence community, including, the National Security Agency's (NSA's) Laboratory for Analytic Science, who has used our work to underpin some of their own research programmes.

We do not yet know what Generation 7 Analytical Tradecraft Training is going to look like. What we do know is that there will be one, and we invite experts across the research community to show us the evidence that will take us to the next level.

5.5 ACKNOWLEDGEMENTS

This chapter was written by analytic tradecraft trainers and researchers who have lived and breathed analytic tradecraft throughout the past decade. Throughout this time there has not been a single dedicated person whose job it was to develop and deliver evidence-based analytic tradecraft training. Everything you have read about here is the work of a small group of dedicated passionate volunteers who have given up their time despite competing pressures. We'd like to thank every single member of our small volunteer army. A final and special thanks to Professor Mandeep K. Dhami who is an inspiration to all of us, and who has been vital in bringing evidence-based policy and tradecraft into our organisation.

5.6 REFERENCES

[1] Butler, F.E.R., Chilcot, J., Inge, P.A., Mates, M., and Taylor, A. (2004). *Review of Intelligence on Weapons of Mass Destruction: Implementation of Its Conclusions.* Retrieved from https://fas.org/irp/world/uk/butler071404.pdf.

[2] Heuer, R.J. (1999). *The Psychology of Intelligence Analysis*. Washington DC: Center for the Study of Intelligence, Central Intelligence Agency.

[3] Belton, I., and Dhami, M.K. (in press). Cognitive biases and debiasing in intelligence analysis. In: *Handbook on Bounded Rationality*, Viale, R., and Katzikopoulos, K. (Eds.). London, UK: Routledge.

[4] Dhami, M.K., Belton, I., and Mandel, D.R. (2019). The 'analysis of competing hypotheses' in intelligence analysis. *Applied Cognitive Psychology*, 33(6):1080-1090.

[5] Mandel, D.R., Karvetski, C.W., and Dhami, M.K. (2018). Boosting intelligence analysts' judgment accuracy: What works, what fails? *Judgment and Decision Making,* 13(6),607-621.

[6] Dhami, M.K., and Careless, K.E. (2015). Ordinal structure of the generic analytic workflow: A survey of intelligence analysts. In: *Proceedings of the 2015 European Intelligence and Security Informatics Conference,* 141-144.

[7] Dhami, M.K., and Careless, K.E. (2019). Intelligence analysts' strategies for solving analytic tasks. *Military Psychology,* 31(2),117-127.

[8] Dhami, M.K., and Careless, K.E. (in press). Development and validation of the analysis support guide. *International Journal of Intelligence and CounterIntelligence*.

[9] Dhami, M.K., Onkal, D., and Wicke, L. (2020). The cone of plausibility. Manuscript submitted for publication.

[10] Wicke, L., Dhami, M.K., Onkal, D., and Belton, I. (2019). Using scenarios to forecast outcomes of the Syrian refugee crisis. *International Journal of Forecasting.* https://doi.org/10.1016/j.ijforecast.2019.05.017.

[11] Dhami, M.K. (2018). Towards an evidence-based approach to communicating uncertainty in intelligence analysis. *Intelligence and National Security* 33(2):257-272.

[12] Ho, E.H., Budescu, D.V., Dhami, M.K., and Mandel, D.R. (2015). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science & Policy,* 1(2):43-55.

[13] Dhami, M.K., and Careless, K.E. (2015). Intelligence analysis: Does collaborative analysis outperform the individual analyst? *The Journal of Intelligence Analysis* 22 (3): 43-58.

[14] Dhami, M.K. (2017). A survey of intelligence analysts' perceptions of analytic tools. In: *Proceedings of the 2017 European Intelligence and Security Informatics Conference,* 131-134.

[15] Dhami, M.K., Belton, I., and Careless, K.E. (2016). Critical review of analytic techniques. In: *Proceedings of the 2016 European Intelligence and Security Informatics Conference,* 152-155.

[16] Phillips, J., Liebowitz, J., and Kisiel, K. (2001). Modeling the intelligence analysis process for intelligent user agent development. *Research and Practice in Human Resource Management,* 9(1):59-73.

[17] Elm, W., Potter, S., Tittle, J., Woods, D., Grossman, J., and Patterson, E. (2005). Finding decision support requirements for effective intelligence analysis tools. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 49(3):297-301.

[18] Pirolli, P., and Card, S.K. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *Proceedings of International Conference on Intelligence Analysis,* 5:2-4.

[19] Klein, G., Phillips, J.K., Rall, E.L., and Peluso, D.A. (2007). A data-frame theory of sensemaking. In: *Expertise Out of Context: Proceedings of the 6th International Conference on Naturalistic Decision Making*, Hoffman, R.R. (Ed.),113-155. Mahwah, NJ: Lawrence Erlbaum Associates.

[20] Moore, D.T. (2011). *Sensemaking: A Structure for an Intelligence Revolution.* Washington, DC: Center for Strategic Intelligence Research, National Defense Intelligence College.

[21] Kang, Y., and Stasko, J. (2011). Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. In: *IEEE Conference on Visual Analytics Science and Technology*. Santucci, G., and Ward, M. (Eds.), 21-30. Piscataway, NJ: IEEE.

[22] UK Ministry of Defence (2013). *Quick Wins for Busy Analysts*, UK.

[23] US Government. (2009). *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis.* Washington, DC: Center for Study of Intelligence Analysis.

[24] Evans, J. St. B.T, and Over, D.E. (1996). *Rationality and Reasoning.* Hove, UK: Psychology Press.

[25] Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist,* 58(9):697-720.

[26] Sloman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin* 119:3-22.

[27] Patterson, E.S., Roth, E.M., and Woods, D.D. (2001). Predicting vulnerabilities in computer-supported inferential analysis under data overload. *Cognition, Technology & Work,* 3(4):224-237.

[28] Pirolli, P., Lee, T., and Card, S.K. (2004). The sensemaking process and leverage points for analyst technology identified through cognitive task analysis. Retrieved from https://www.e-education.psu.edu/geog885/sites/www.e-education.psu.edu.geog885/files/geog885q/file/Lesson_02/Sense_Making_206_Camera_Ready_Paper.pdf

[29] Trent, S.A., Patterson, E.S., and Woods, D.D. (2007). *Team cognition in intelligence analysis*. Proceedings of Human Factors and Ergonomics Society Annual Meeting, 51(4),308-312.

[30] Chin, G., Jr., Kuchar, O.A., and Wolf, K.E. (2009). Exploring the analytical processes of intelligence analysts. In: *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pp. 11-20.

[31] Roth, E.M., Pfautz, J.D., Mahoney, S.M., Powell, G.M., Carlson, E.C., Guarino, S.L., Fichtl, T.C., and Potter, S.S. (2010). Framing and contextualizing information requests: Problem formation as part of the intelligence analysis process. *Journal of Cognitive Engineering and Decision Making,* 4(3):210-239.

[32] Dhami, M.K., Belton, I., and Goodman-Delahunty, J. (2015). Quasi-rational models sentencing. *Journal of Applied Research on Memory and Cognition,* 4(3):239-247.

[33] Dhami, M.K., and Careless, K.E. (2020). A test of the 'analysis support guide'. Manuscript in preparation.

[34] Dhami, M.K., Mandel, D.R., Mellers, B.A., and Tetlock, P.E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science,* 10(6):753-757.



Part II: INFORMATION EVALUATION UNDER UNCERTAINTY



Chapter 6 – APPLYING INFORMATION THEORY TO VALIDATE COMMANDERS' CRITICAL INFORMATION REQUIREMENTS[1]

Mark A. C. Timms
Department of National Defence
CANADA

David R. Mandel
Defence Research and Development Canada
CANADA

Jonathan D. Nelson
University of Surrey
UNITED KINGDOM

6.1 INTRODUCTION

The primary aim of this chapter is to introduce a novel approach to strengthen contemporary intelligence community practices for establishing intelligence collection priorities based on expected information value. We propose the integration of quantitative measures of information utility that have been discussed in the literature on information theory [2], [3], [4] as a method for optimizing intelligence collection planning. We argue that enhancing the effectiveness through which command information requirements are established can improve consequent intelligence collection priorities. We contrast this approach with the Structured Analytic Technique (SAT) approach that is currently described as a method for prioritizing information requirements in intelligence collection. Specifically, we proceed with a review of the Indicators Validator™ (IV) SAT [5] for establishing information value, illustrating how it works, and where it falls short as an analytic method. Next, we introduce a quantitative information-theoretic measure of information utility called *information gain* [1]. We illustrate the contrast between these approaches using a practical example featuring a hypothetical North Atlantic Treaty Organization (NATO) dilemma. This analysis shows how information gain overcomes many limitations of the IV technique, along with how it might be applied to modern NATO operational practice.

6.2 BACKGROUND: THE NATO INTELLIGENCE COMMUNITY DILEMMA

Intelligence organizations iteratively explore new ways to assess information value. NATO intelligence professionals inform complex, high-consequence, operational decisions on a routine basis. First established in 1949, NATO's stated purpose is to "…guarantee the freedom and security of its members through political and military means" [6]. Where a NATO force has been deployed to monitor another government's adherence to cease-fire agreements, the success of its mandate could become entirely dependent on its ability to accurately interpret indicators of imminent aggression. Under these circumstances, failing to act when required can be just as damaging as taking action when none is warranted. Intuitively, when charged with such a fragile task, a NATO commander would want to position his forces in such a way that allow the initiation of swift, deliberate, and decisive intervention (if and when required), without adopting a force posture that inadvertently encourages or re-ignites existing tensions between hostile states.

The NATO Intelligence Community (IC) enhances command understanding of complex environments through the delivery of predictive assessments founded in the deliberate analysis of threat event indicators and warning

[1] Correspondence concerning this chapter should be sent to David Mandel at David.Mandel@drdc-rddc.gc.ca. This chapter is reprinted with permission from the publisher from the following source: Timms et al. [1]. We thank two anonymous reviewers and the editors of that volume for their feedback on earlier drafts of this chapter. This work was funded by Canadian Safety and Security Program Projects CSSP-2016-TI-2224 and CSSP-2018-TI-2394 under the scientific direction of David Mandel.

signs. Whether the analysis is intended to provide context or early warning, or to identify opportunities, it is fundamentally about improving decision making under conditions of uncertainty (see Ref. [7], p. 5). Once a mission or political mandate is defined, intelligence professionals are often left to identify which questions, if answered, can most efficiently improve stakeholder decision making in the context of that mission. We suggest that the consistent, coherent, and precise evaluation of information usefulness during the earliest stages of operational planning is vital to ensuring sound intelligence collection planning, although some literature suggests that members of the operational community often only pay it lip service to this aim [8].

6.2.1 Establishing Command Information Priorities

In order to prioritize organizational resourcing, decision makers issue a series of Information Requirements (IR) to subordinate units that subsequently drive intelligence collection efforts. Some of those IR (i.e., questions), when answered, may compel stakeholders to take action, cease action, or have some form of immediate impact on organizational posture. These are often called Commanders' Critical Information Requirements (CCIR) (see Ref. [9], pp. 3-2). Although commanders issue planning guidance, individual planners (or groups) often develop CCIR through subjective, multi-stage planning processes that vary across organizations. CCIR can drive the assignment of collection resources to fill information gaps (see Ref. [10], pp. 1-2), and they are often presented to non-expert decision makers for approval without having been formally evaluated to ensure that their answers will actually reduce uncertainty in command decision making in some appreciable manner [8].

Figure 6-1 depicts the hierarchical relationship between Commanders' Critical Information Requirements (CCIR) and related information priorities developed through consequent processes, including Priority Intelligence Requirements (PIR) and Friendly Force Information Requirements (FFIR), each of which drive the defining of Essential Enemy Information (EEI), Essential Elements of Friendly Information (EEFI) and more (adapted from Ref. [8]). Numerous agencies may define CCIR (and other consequent IRs) through subjective group (or individual) brainstorming. The success of such exercises will likely depend on the abilities of the individual planner (or planning group), varying in terms of the diversity, size, and breadth of experience (among other items) of the members involved. At the very least, planners must be capable of shaping proposed CCIR from their translation of such intent, through formulating plans to meet those requirements.

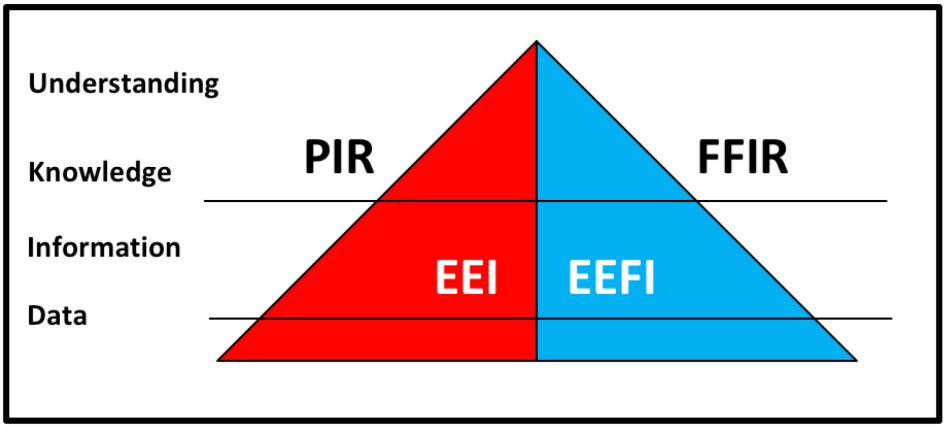


Figure 6-1: Hierarchy of Information Requirements.

6.3 THE SAT APPROACH

The IV technique that we consider later in this chapter is an example of the SAT approach adopted over several decades within many NATO countries. The development of SATs started out as a rather idiosyncratic,

path-dependent endeavor spearheaded by US intelligence tradecraft mavericks, such as Richards Heuer Jr. and Jack Davis, who took it upon themselves while working at the Central Intelligence Agency (CIA) to develop back-of-the-napkin techniques to aid intelligence analysts [11]. The impetus for developing such methods was, in large part, the belief that analysts were prone to cognitive biases that SATs would help effectively overcome. In this view, SATs could be used to structure the otherwise unbridled intuitions of analysts and to tame their purported wanton subjectivity. The effort to develop SATs and require that analysts be trained to use them received intermittent commitment within the US IC until pivotal geopolitical events – namely, the 9/11 terrorist attacks against the US by Al Qaeda and the faulty, invasion-prompting intelligence estimate that Saddam Hussein was developing weapons of mass destruction in Iraq – triggered congressionally mandated institutional reforms that required the use of SATs by the US IC [10], [12], [13], [14], [15]. The CIA's tradecraft manual originally included 12 SATs [16], but the list of SATs has burgeoned, now including several dozen such techniques [5], [17].

The body of scientific research on SATs (and analytic tradecraft, more generally) remains scant [13], [18], [19], [20]. Unfortunately, the IC has tended to view SATs as benign if not beneficial. While SAT proponents will often admit that SATs "aren't perfect", they are usually quick to add that they are "better than nothing." However, recent evidence indicates that the latter supposition may be false. Mandel, Karvetski, and Dhami [21] studied the effects of training analysts in the use of one particular SAT, the Analysis of Competing Hypotheses [17], [22]. Remarkably, analysts who used that SAT to assess the probability of alternative hypotheses were significantly *less* coherent and also less accurate in their judgements than analysts who were not instructed to use any SAT (also see Ref. [23]). Such findings should not be unexpected given that SATs, more generally, are subject to two important conceptual shortcomings [11], [13]. First, they neglect the fact that most cognitive biases are bipolar (e.g., calibrated confidence is offset by underconfidence *or* overconfidence) and they fail to assess the types of biases analysts are in fact prone to before intervening. Second, SATs neglect the cost of noisy judgements that follow from techniques that, though supposedly objective, in fact invite a range of implementation-related decisions that are left to analysts' discretion. Thus, SATs may do more to redirect subjectivity from substantive assessment to resolving methodological vagueness.

Few SATs focus explicitly on evaluating information utility. Those that do are geared towards establishing the predictive value of threat event indicators in the context of impending hypothetical threat events; namely, the Indicators (see Refs. [5]; [17], p. 149) and Indicators Validator™ (IV) (see Refs. [5]; [17], p. 157) SATs. Indicators are defined as: "…observable phenomena that can be periodically reviewed to help track events, spot emerging trends, and warn of anticipated changes" (see Refs. [5]; [17], p. 149). The Indicators SAT encourages analysts to leverage their personal experience in concert with easily accessible information in the development of a detailed indicators list. This list reflects a "pre-established set of observable or potentially observable actions, conditions, facts, or events whose simultaneous occurrence would strongly argue that a phenomenon is present, or at least highly likely to occur" (see Refs. [5]; [17], p. 149). Heuer and Pherson's Indicators SAT encourages analysts to build a list of indicators presumed to be associated with hypothetical threat events.

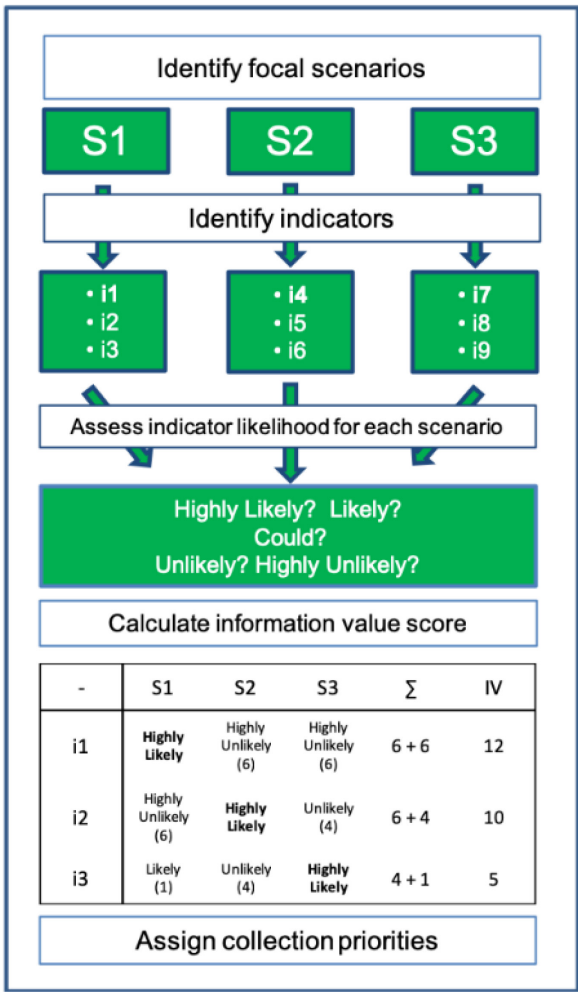
For instance, if you were wondering whether it was going to rain, you might reflexively consider checking the ambient barometric pressure, listening for thunder, or perhaps looking for lightning. Heuer and Pherson's [17] Indicators SAT suggests that NATO intelligence professionals perform similar exercises when attempting to predict events of operational interest. Hypothetical examples include (but are not limited to): whether a political official will be re-elected before cease-fire agreements are signed, the volume of refugees that might move down a series of different corridors in the aftermath of a natural disaster, or whether a small military force might suddenly annex a sovereign bordering state. Whereas the Indicators SAT focuses on indicator definition, its companion IV SAT aims to assist analysts in establishing the predictive value of indicators in the context of a given threat event scenario. In the next section, we review the IV SAT and analyze its performance. Later, to illustrate differences between the IV SAT and more formal models of information utility, we introduce a relevant example inspired by NATO's Enhanced Forward Presence (EFP) forces stationed in Eastern Europe.

6.4 THE INDICATORS VALIDATOR™ SAT

Introduced by Heuer and Pherson in 2008, the IV SAT focuses on establishing the predictive value of an indicator based on how exclusively it indicates a focal hypothesis or threat scenario among a set of scenarios [5], [17]. According to Heuer and Pherson, indicators are "…observable phenomena that can be periodically reviewed to help track events, spot emerging trends, and warn of anticipated changes" (see Ref. [17], 149).

To use the IV SAT, analysts must first identify a list of mutually exclusive and collectively exhaustive threat scenarios (sometimes called hypotheses) to be predicted. These can be multi-alternative (Event A: Person X is elected, Event B: Person Y is elected, Event C: Person Z is elected) or binary (yes/no, happened/did not happen). Each scenario is accompanied by a list of primary indicators that analysts believe would be likely to be present if that scenario were to occur (or if the hypothesis were true). Indicators that are generated for a particular scenario are said to be "at home" for that scenario and are not "at home" for the alternative scenarios. That is, any given indicator can only be "at home" in one scenario for the IV SAT to work as intended. However, any given scenario may have multiple indicators that are "at home" in it.

After assigning indicators to their home scenarios, the analyst must judge whether each indicator is to be rated as *likely* or *highly likely* in its home scenario, as this will affect the consequent information value scoring procedure (see Figure 6-2).



-	S1	S2	S3	∑	IV
i1	**Highly Likely**	Highly Unlikely (6)	Highly Unlikely (6)	6 + 6	12
i2	Highly Unlikely (6)	**Highly Likely**	Unlikely (4)	6 + 4	10
i3	Likely (1)	Unlikely (4)	**Highly Likely**	4 + 1	5

Figure 6-2: The Indicators Validator™ Model (see Ref. [16], 157).

At this stage, the analyst cannot select other probability values (e.g., *very unlikely*) to represent the indicator. In other words, an indicator that is "at home" must be judged to be either *likely* or *very likely*, given that scenario. Next, the likelihood of each indicator given each of the alternative scenarios is assessed. For example, if an indicator is deemed to be *highly likely* given the home scenario, then numerical values would be assigned to the indicator in the alternative scenarios as a function of how divergent they are from the original "at home" rating using the following coding scheme: *highly likely* = 0 (i.e., no divergence), *likely* = 1, *could* = 2, *unlikely* = 4 *highly unlikely* = 6 (i.e., maximum divergence) (see Ref. [17], 159). The IV SAT makes an adjustment for whether the indicator is *highly likely* or only *likely* in the home scenario; if it is judged to be likely, a similar rating scheme is applied but the "distance scores" are smaller in magnitude (see Ref. [17], 159). Finally, the analyst would sum the distance scores assigned to the alternative scenarios. The greater the summed distance score, the more useful a given indicator is deemed to be.

6.4.1 Analysis of the IV SAT

The IV SAT is a simple method for scoring the usefulness of different indicators. The ordering of the steps in the technique is easy to follow and the application of the IV SAT does not require mathematical sophistication. Nevertheless, in spite of its ease of application, the IV SAT has important limitations that could lead collectors astray, and misinform analysts and intelligence consumers.

One problem with the technique is the arbitrariness of matching indicators to home scenarios. For an indicator to be "at home" it must be judged either *likely* or *very unlikely* given the scenario. However, it may be judged equally likely under other scenarios, in which case the indicator might just as well have been "at home" in those scenarios. This aspect of the technique highlights another sense in which it is arbitrary – namely, it disallows negative hypothesis tests in which one searches for indicators that may be (*very*) *unlikely* if the scenario were true. Detecting the presence of such low probability events can be highly informative, yet the IV SAT precludes such a focus in prioritization of information requirements. One remarkable implication of this constraint is that the complement of a good indicator (i.e., one that has low probability in the home scenario but high probability in all of the other scenarios) would be precluded from being considered. This is not only arbitrary; it is incoherent.

Another limitation of the IV SAT is that it does not require or even prompt analysts to consider the prior probability of scenarios or indicators. It does not do so either in terms of collection of objective, relative frequency data that could be used to establish base-rate estimates or in terms of subjective estimates of these relative frequencies, which could also be useful. This critical base-rate information is accounted for in virtually all information-theoretic models [1], [3], [24], and also in Bayesian approaches to belief revision [25]. Collection resources are not unlimited [26], necessitating careful evaluation and IR prioritization. We suggest that it would be useful to first establish how likely a commander is to correctly predict an outcome based on what is already known about event base rates, and then to evaluate the utility of information in the context of how much it improves event predictability over reliance on prior probabilities alone. If that which is already known about a given threat event can enable a commander to confidently predict its occurrence, collection assets might more efficiently be directed towards answering questions that would measurably reduce uncertainty associated with other events.

Consider an indicator that would almost certainly signal the occurrence of Scenario 1 (Attack) – such as an intercepted Russian correspondence directing military forces to cross their border at a given time – where the actual probability of an attack occurring was assessed to be extremely low. The IV SAT would award a high information value score to that indicator, plausibly resulting in it becoming a priority collection item, despite the fact that its likelihood of appearance is significantly less likely than competing indicators in other scenarios, because it is "at home" in an improbable scenario.

The IV technique is the only information evaluation SAT featured in open-source intelligence analytic tradecraft manuals for both Canadian military [7] and American [8] intelligence agencies. Unfortunately, for the reasons noted, it lacks a sound, logical foundation. However, its vague quantification of individual probability judgements using linguistic probabilities on an ordinal scale, and the procedures used for calculating a final information value score for an indicator, can lead analysts to believe that they are following a valid, even objective, method.

The IV SAT assigns indicator value as a function of how strongly its presence predicts a single threat scenario. The final information value score focuses only on the extent to which the indicator can discriminate between the scenario in which it is "at home" and alternatives in the same set. This could dramatically reduce efficiency in collection planning. In our example, for simplicity, we limit the number of threat scenarios to three. But in many cases, there will be multiple threat scenarios of interest that are thematically similar but ultimately distinct. In such cases, a low information value score for an indicator could be deceiving. Furthermore, consider an indicator that is strongly associated with the occurrence of all but one scenario. IV would give such an indicator a low information value score, despite the fact that it could reliably help a NATO commander predict the event in which it is not present.

With this in mind, we present an alternative approach to evaluating and prioritizing command information requirements, using information gain [1], [3]. In many ways, the structure of information-theoretic measures compels increased analytic reasoning, as the various inputs of the information gain formula may require the conduct of research, or the deliberate assignment of a numeric probability to a threat event scenario or its co-occurrence frequency with indicators. Importantly, information gain requires some input values for the probability of each scenario. The need to include estimates about each scenario's probability can encourage analysts to reduce the uncertainty associated with an entire problem, rather than pursuing information associated with events that may already be relatively easy to predict.

6.5 INFORMATION GAIN: A PRINCIPLED APPROACH TO EVALUATING INDICATOR USEFULNESS

In this section, we describe a quantitative information-theoretic measure of information utility called *information gain* that measures the average reduction in uncertainty achieved by using a specific indicator or cue [1], [3]. Information gain is an example of an information utility function, a mathematical formula designed to compute a quantitative estimate of utility for a piece of information. Superficially, information utility functions, like information gain, require similar inputs to those required when using the IV SAT. The basic principle remains: first, define information gaps (i.e., what one wants to know); next, identify what questions (and answers) might help fill them. The expected information value of a question is ultimately defined as the expected value of the not-yet-obtained answer, although the value of specific answer could also be calculated [3], [27]. In contrast to the IV SAT, information gain has been effectively used in a variety of domains, such as automatic face recognition systems [28], image registration [29], predicting human queries [30], philosophy of science [4], and modeling neurons in visual [31], [32] and auditory perception [33].

6.5.1 Computing Information Gain

Information gain quantifies the utility of a given indicator as a function of how effectively its presence or absence reduces uncertainty about a hypothetical event of interest [3]. Lindley [1], Box and Hill [34], and Fedorov [35] quantified this idea explicitly, using Shannon's [36] entropy to measure the uncertainty in the outcome of a specific event. We measure information with base 2 logarithms (bits). Other bases could also be used. If the natural logarithm is used, the unit is *nats*.

For the purposes of defining information gain, let $Q = \{q_1, q_2, \ldots q_m\}$ represent a query (in mathematical terms, a random variable), in this case, the option of querying the value of a particular threat event indicator or

command information requirement. Let each q_j represent one of the m possible answers to the question Q. Let $H = \{h_1, h_2 ... h_n\}$ represent the unknown hypothesis (or category or threat scenario) one is trying to predict. Finally, let each of the n possible h_i represent a specific hypothesis in the set of possibilities (i.e., a list of mutually exclusive and exhaustive threat event scenarios). Equation 1 shows the information gain calculation:

$$I(H,Q) = \left[\sum_{i=1}^{n} P(h_i) * \log_2 \frac{1}{P(h_i)}\right] - \left[\sum_{q=1}^{m} P(q_j) * \sum_{i=1}^{n} * P(h_i|q_j) * \log_2 \frac{1}{P(h_i|q_j)}\right] \quad (6\text{-}1)$$

Information gain for a given indicator is equal to the initial entropy minus the entropy that is expected (on average) to be remaining after the indicator's state (e.g., present or absent) is observed. In other words, information gain measures the change in Shannon entropy from before (i.e., base-rate scenario uncertainty) to after consideration of the indicator's state [3]. Information gain can be used with indicators having two or more possible states. If information gain were used to prioritize collection, then the indicator with greatest expected reduction in uncertainty across the whole set of threat scenarios would be rated as the top priority for subsequent collection activities. Information gain is also known as the mutual information [37] between the hypotheses of interest H and the indicator Q.

6.6 APPLYING IV AND INFORMATION GAIN TO THE NATO EXAMPLE

Currently, there are numerous battalion-sized (300 – 1300 soldiers) military units from contributing NATO member nations occupying a defence and deterrence posture in several countries along the Russian border [38]. Hypothetically, if one of these units intercepted correspondence that Russia intended to conduct a large-scale training event in the near future, this might trigger the formation of an incident-based planning group, where available staff officers would convene and think through new information with a view to presenting their commander with options for implementation. Imagine that a planning group is convened. The group must generate a prioritized list of information requirements associated with the Russian exercise, with a view to helping their commander determine their force posture during the exercise, whether reinforcements will be required, and more.

Using the IV SAT, the group would first flesh out a list of mutually exclusive and collectively exhaustive hypotheses, each of which might compel their commander to take or delay a specific action (kept relatively simple here, see Table 6-1). Scenario 1, from the infantry: the Russians are staging for an attack. The infantry planner also proposes her top indicator for this scenario: live ammunition. The idea is that if the Russians were staging for an attack, they would most certainly be carrying live ammunition. Scenario 2, from the logistician: The Russians intend to carry out a training exercise, sincerely aiming to improve the quality and professionalism of their forces through the practice of large military maneuvers. The logistician highlights that soldiers feed differently under combat conditions than they do in training. Large-scale training events are likely to implicate the use of a non-tactical field feeding kitchen system for soldiers participating in training. Finally, the public affairs officer proposes another possibility, Scenario 3, namely that the Russians are actually posturing, conducting a show of force to NATO, to communicate that the multinational posture has not impacted their resolve. The public affairs officer further suggests that, if this scenario were to occur, the Russians would communicate their message through deliberate media events, such as press conferences.

Table 6-1: IV Matrix for the Scenario. Estimates for "at home" indicators are bolded.

Indicators	S1: Attack	S2: Exercise	S3: Posturing	Information Value Score	Collection Priority
I1: Live Ammunition	**Highly Likely**	Highly Unlikely (6)	Highly Unlikely (6)	12	1
I2: Field Kitchen	Highly Unlikely (6)	**Highly Likely**	Unlikely (4)	10	2

I3: Media Events	Likely (1)	Unlikely (4)	**Highly Likely**	5	3

Next, planners debate indicator/scenario co-occurrence frequencies. In Table 6-1, I1: Live Ammunition is "at home" in S1: Attack. The planners judge that Russian soldiers are *highly likely* to be carrying live ammunition when they stage before a combat event. I1 is agreed to be *highly unlikely* to be present in the other two scenarios, which earns it an IV information value score of 12, thus moving it to top priority for collection assets, followed closely by I2 (IV score: 10), with I3 well behind its companions (IV score: 5). Thus, the IV SAT prioritizes the indicators as follows: I1 > I2 > I3. Because these numbers are generated on an ordinal scale, differences in information value score do not directly reveal proportional increases (i.e., the fact that I3 = 5, and I2 = 10 does not mean that I3 is half as valuable as I2).

Next, we consider how information gain might be applied in this scenario. In Table 6-2 we have included additional planner estimates of event base rates for each of the threat event scenarios, where the probability (*P*) of S1, a deliberate military attack, is considered low (1%); the other scenarios are deemed much more likely to occur, with the probability of an exercise (S2) at a 33% chance, and the probability of posturing (S3) at a 66% chance. Table 6-2 illustrates that the information value scores applied to the same indicators using information gain produce the opposite prioritization as the IV SAT. That is, using information gain, the expected values (in bits) of the indicators are: I1 = 0.0249, I2 = 0.0408, and I3 = 0.2722. This results in a collection priority assignment of I3 > I2 > I1. Clearly, the choice of method used to evaluate information usefulness can have dramatic consequences, including the full reversal of recommended collection priorities.

Table 6-2: Information Gain Assessment for the Scenario.

Indicators	**S1: Attack**	**S2: Exercise**	**S3: Posturing**	**Information Value Score**	**Collection Priority**
Prior Probabilities	**0.01**	**0.33**	**0.66**	**-**	**-**
I1: Live Ammunition	0.9	0.1	0.1	0.0249	3
I2: Field Kitchen	0.1	0.9	0.75	0.0408	2
I3: Media Events	0.75	0.3	0.9	0.2721	1

6.7 DISCUSSION

As demonstrated in our example scenario in Table 6-1 and Table 6-2, both the IV SAT and information gain can be used to facilitate prioritization of IR in support of information collection and decision making. The IV SAT generates a rank-ordered list of indicator usefulness. Information gain provides a continuous ratio measure of the expected information value of alternative indicators. Although both the IV SAT and information gain require analysts to assess probabilities, the IV SAT imposes arbitrary constraints on what probabilities may be applied. Indeed, for the focal ("home") hypothesis, only two possibilities are permitted: likely or highly likely. In contrast, information gain does not impose arbitrary rules on probability assignment. Moreover, unlike the IV SAT, information-theoretic measures such as information gain do not impose coarseness on the probabilities assigned. Recent research has shown that imposing coarseness on more granular probability judgements reduces accuracy across a wide range of conditions [39]. Although analysts may initially balk at the idea of providing more granular judgements, Barnes [40] found that they rapidly adjust to making granular assessments and are willing to debate about differences that would otherwise have been obscured. The use of information-theoretic measures is therefore in step with recent recommendations to use numeric probabilities in intelligence production [18], [39], [40], [41].

Strikingly, our hypothetical NATO example shows that the IV SAT can deliver information collection priorities in total opposition to those generated by information gain. Given that information gain (in contrast to the IV SAT) has proven itself robust and useful over many decades in diverse contexts, this striking discrepancy should be disconcerting to operational communities that rely on the IV SAT or something like it. Why do the IV SAT and information gain contradict each other in this example? A key reason is that information gain accounts for the base rate of each threat scenario when computing information value. By comparison, the IV SAT is fully insensitive to these base rates. In this sense, the IV SAT formalizes, and perhaps even reinforces, base-rate neglect [42], [43], whereas information-theoretic measures like information gain should mitigate this form of bias.

Information-theoretic metrics such as information gain could be tested through future research, completed in collaboration with military intelligence professionals, or anyone seeking to strengthen the manner in which they establish and validate command information requirements, on and off the battlefield. Information gain, based on expected reduction in Shannon entropy across the set of all possible hypotheses, is a widely used method. However, expected reduction in other kinds of entropy measures or qualitatively different information utility functions could also be used. The differences between information gain and related information-theoretic approaches are in many cases not dramatic [3], in contrast with the differences between information gain and the IV SAT. For conciseness and because of its robustness and wide use in many domains, we have based the numeric example in this chapter on information gain. Nelson (see Ref. [3], Appendix A) provides example numeric calculations, and Crupi *et al*. [30] show how many different entropy and information measures from mathematics and physics can be articulated within a unified formal framework.

In future work, it would be worthwhile to go beyond our toy example and to consider actual relevant scenarios using these methods, considering also which of many possible methods are most appropriate. It would also be useful to consider the implications of including extra-informational factors, such as the cost of acquiring specific pieces of information, and potentially asymmetric costs of different kinds of mistakes, such as false-positive vs. false-negative errors, when evaluating the expected usefulness of possible indicators. A further qualification is that in our scenario, we consider the usefulness of each indicator individually. Depending on the dependency structure in the domain, if more than one indicator can potentially be queried, it may be necessary to evaluate the information value of possible sequences (decision trees) of indicator evaluation. The information-theoretic approaches generalize naturally to situations with known dependencies among indicators (see Ref. [44]).

Since access to quantitative, frequentist data for intelligence analysis is often lacking [45], future research might also explore optimal methods for compiling, organizing, and structuring such intelligence data. Additionally, since there is currently considerable apprehension in the intelligence community to using quantitative methods that require at least a rudimentary understanding of statistics and probability, intelligence management would have to better train analysts in these respects. Even brief training in probabilistic reasoning using natural frequency formats has been shown to improve analysts' logical coherence and accuracy [46]. Intelligence organizations might do well to de-emphasize some training that is of questionable value, such as current SAT tradecraft training [13], [19]. Rather than introducing new SATs that may have no more than face validity, it would be useful – as in the case of evaluating the expected usefulness of potential information sources – to consider whether other disciplines have robust techniques that could be adopted in defence and security contexts. As noted elsewhere [11], [18], [19], [20], in support of that objective, the defence and security community should exploit the interdisciplinary decision sciences.

6.8 REFERENCES

[1] Timms, M., Mandel, D.R., and Nelson, J.D. (2020). Applying information theory to validate commanders' critical information requirements. In: *Handbook of Military and Defence Operation Research*, Scala, N.M., and Howard, J. (Eds.), 331-344. Boca Raton, FL: CRC Press.

[2] Lindley, D.V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics,* 27(4):986-1005.

[3] Nelson, J.D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review* 112 (4):979-999.

[4] Crupi, V., and Tentori, K. (2014). State of the field: Measuring information and confirmation. *Studies in History and Philosophy of Science Part A* 47:81-90.

[5] Heuer, R.J., Jr., and Pherson, R.H. (2008). *Structured Analytic Techniques for Intelligence Analysis*. Washington DC: CQ Press. Developed by Pherson Associates, LLC.

[6] North Atlantic Treaty Organization. (2017). Website homepage. Retrieved from http://www.nato.int/nato-welcome/index.html.

[7] Canadian Forces Intelligence Command. (2016). *Analytic Writing Guide* (Version 3.0). Department of National Defence: Ottawa, Canada.

[8] US Government. (2013). *Insights and Best Practices Focus Paper: Commander's Critical Information Requirements. Deployable Training Division (DTD),* Deputy Director Joint Staff J7, Joint Training. Retrieved from http://www.dtic.mil/doctrine/fp/fp_ccirs.pdf.

[9] Commander Canadian Army. (2013). *Intelligence, Surveillance, Target Acquisition, and Reconnaissance (ISTAR) Volume 1 – The Enduring Doctrine*. B-GL-352-001/FP-001. Department of National Defence: Ottawa, Canada.

[10] Chief of Defence Staff. (2002). *CF Operational Planning Process*. B-GJ-005-500/FP-000. Department of National Defence: Ottawa, Canada.

[11] Mandel, D.R., and Tetlock, P.E. (2018). Correcting judgment correctives in national security intelligence. *Frontiers in Psychology,* 9:2640.

[12] Artner, S., Girven, R.S., and Bruce, J.B. (2016). *Assessing the Value of Structured Analytic Techniques in the US Intelligence Community*. Santa Monica, CA: RAND Corporation.

[13] Chang, W., Berdini, E., Mandel, D.R., and Tetlock, P.E. (2018). Restructuring our thinking about structured techniques in intelligence analysis. *Intelligence and National Security* 33 (3):337-356.

[14] Coulthart, S.J. (2017). An evidence-based evaluation of 12 core structured analytic techniques. *International Journal of Intelligence and Counter Intelligence,* 30(2):368-391.

[15] Marchio, J. (2014). Analytic tradecraft and the intelligence community: Enduring value, intermittent emphasis. *Intelligence and National Security* 29 (2):159-183.

[16] US Government. (2009). *Structured Analytic Techniques for Improving Intelligence Analysis*. Washington DC: Central Intelligence Agency, Center for the Study of Intelligence.

[17] Heuer, R.J., Jr., and Pherson, R.H. (2014). *Structured Analytic Techniques for Intelligence Analysis.* Washington DC: CQ Press. Developed by Pherson Associates, LLC.

[18] Dhami, M.K., Mandel, D.R., Mellers, B.A., and Tetlock, P.E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science,* 10(6):753-757.

[19] Mandel, D.R. (2019). Can decision science improve intelligence analysis? In: *Researching National Security Intelligence: Multidisciplinary Approaches*, Coulthart, S., Landon-Murray, M., and Van Puyvelde, D. (Eds.), 117-140. Washington DC: Georgetown University Press.

[20] Pool, R. (2010). *Field Evaluation in the Intelligence and Counterintelligence Context: Workshop Summary*. Washington DC: The National Academies Press.

[21] Mandel, D.R., Karvetski, C.W., and Dhami, M.K. (2018). Boosting intelligence analysts' judgment accuracy: What works, what fails? *Judgment and Decision Making* 13 (6):607-621.

[22] Heuer, R.J., Jr. (1999). *The Psychology of Intelligence Analysis*. Washington DC: Central Intelligence Agency, Center for the Study of Intelligence.

[23] Dhami, M.K., Belton, I.K., and Mandel, D.R. (2019). The "analysis of competing hypotheses" in intelligence analysis. *Applied Cognitive Psychology*, 33(6):1080-1090.

[24] Wu, C.M., Meder, B., Filimon, F., and Nelson, J.D. (2017). Asking better questions: How presentation formats influence information search. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 43(8):1274-1297.

[25] Navarrete, G., and Mandel, D.R. (Eds.). (2016). *Improving Bayesian Reasoning: What Works and Why?* Lausanne, Switzerland: Frontiers Media.

[26] Folker, R.D., Jr. (2000). *Intelligence Analysis in theater joint intelligence centers: An experiment in applying structured methods*. Washington DC: Center for Strategic Intelligence Research, National Defense Intelligence College.

[27] Nelson, J.D. (2008). Towards a rational theory of human information acquisition. In: *The Probabilistic Mind: Prospects for Rational Models of Cognition*, Oaksford, M., and Chater, N. (Eds.), 143-163. Oxford, UK: Oxford University Press.

[28] Imaoka, H., and Okajima, K. (2004). An algorithm for the detection of faces on the basis of Gabor features and information maximization. *Neural Computation,* 16(6):1163-1191.

[29] Chen, H.M., Arora, M.K., and Varshney, P.K. (2003). Mutual information-based image registration for remote sensing data. International *Journal of Remote Sensing* 24 (18):3701-3706.

[30] Crupi, V., Nelson, J.D., Meder, B. Cevolani, G., and Tentori, K. (2018). Generalized information theory meets human cognition: Introducing a unified framework to model uncertainty and information search. *Cognitive Science,* 42(5):1410-1456.

[31] Ruderman, D.L. (1994). Designing receptive fields for highest fidelity. *Network: Computation in Neural Systems,* 5(2):147-155.

[32] Ullman, S., Vidal-Naquet, M., and Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience,* 5(7):682-687.

[33] Lewicki, M.S. (2002). Efficient coding of natural sounds. *Nature Neuroscience,* 5 (4):356-363.

[34] Box, G., and Hill, W. (1967). Discrimination among mechanistic models. *Technometrics,* 9(1):57-71.

[35] Fedorov, V.V. (1972). *Theory of Optimal Experiments*. New York, NY: Academic Press.

[36] Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379-423.

[37] Cover, T.M., and Thomas, J.A. (2012). *Elements of Information Theory*. New York, NY: John Wiley & Sons.

[38] North Atlantic Treaty Organization. (2017). Boosting NATO's presence in the east and southeast. Retrieved from: https://www.nato.int/cps/en/natohq/topics_136388.htm.

[39] Friedman, J. (2019). *War and Chance: Assessing Uncertainty in International Politics.* New York, NY: Oxford University Press.

[40] Barnes, A. (2016). Making intelligence analysis more intelligent: Using numeric probabilities. *Intelligence and National Security,* 31(3):327-344.

[41] Irwin, D., and Mandel, D.R. (2019). Improving information evaluation or intelligence production. *Intelligence and National Security,* 34(4):503-525.

[42] Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica,* 44(3):211-233.

[43] Kahneman, D., and Tversky, A. (1973). On the psychology of prediction. *Psychological Review,* 80(4):237-251.

[44] Nelson, J.D., Meder, B., and Jones., M. (2018). Towards a theory of heuristic and optimal planning for sequential information search.

[45] Spielmann, K. (2016). I got algorithm: Can there be a Nate Silver in intelligence? *International Journal of Intelligence and CounterIntelligence* 29 (3):525-544.

[46] Mandel, D.R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Frontiers in Psychology,* 6:387.

Chapter 7 – STANDARDS FOR EVALUATING SOURCE RELIABILITY AND INFORMATION CREDIBILITY IN INTELLIGENCE PRODUCTION[1]

Daniel Irwin and David R. Mandel
Defence Research and Development Canada
CANADA

7.1 INTRODUCTION

Intelligence practitioners must regularly exploit information of uncertain quality to support decision making [1]. Whether information is obtained from a human source or an automated sensor, failure to assess and communicate its characteristics may contribute to intelligence failure [2], [3]. This is evident in the case of Curveball, the Iraqi informant who fabricated extensive testimony on Saddam Hussein's alleged Weapons of Mass Destruction (WMD) [4], [5]. Subjected to inadequate scrutiny, Curveball's false allegations underpinned the 2002 National Intelligence Estimate on Iraq's WMD programs, and may have influenced the ill-fated decision to invade Iraq in 2003 [4], [5].

Recognizing information evaluation as a key function within the intelligence process, some organizations provide standards for assessing and communicating relevant information characteristics. Despite their intent, however, many of these standards are inconsistent across organizations, and may be fundamentally flawed or otherwise ill-suited to the context of application. In certain situations, poorly formulated standards may actually inhibit collaboration, degrade the quality of analytic judgements, and impair decision making.

In order to develop evidence-based recommendations for future practice in the assessment and communication of information quality, SAS-114 collected standards in use across a variety of agencies and domains. The following chapter provides a critical examination of standards for evaluating source reliability and information credibility, and highlights avenues for future research and development.

7.2 OVERVIEW OF CURRENT STANDARDS

The information evaluation criteria presented in Allied intelligence doctrine is known as the Admiralty Code or NATO System [6]. Developed by the Royal Navy in the 1940s, the system has undergone little change since its inception, and forms the basis of standards used by several Alliance members, as well as organizations in other domains [7], [8]. Under the Admiralty Code, information is assessed on two dimensions: source reliability and information credibility. Users are instructed to consider these components independently and to rate them on two separate scales (Table 7-1). The resultant rating is expressed using the corresponding alphanumeric code (e.g., *probably true* information from a *usually reliable* source is rated B2). Both scales include an option to be used when there is an inability to assess ('F' for source reliability and '6' for information credibility). Thus, ratings 'F' and '6' are not part of the ordinal scales comprised of ratings A – E and 1 – 5, respectively.

The defunct NATO Standardization Agreement (STANAG) 2511 (superseded in terms of information evaluation standards by Ref. [9]) provides a more detailed version of the Admiralty Code, and is presented for historical reference in Table 7-2 and Table 7-3 [10]. In line with many of the standards examined, NATO STANAG 2511 includes a qualitative description for each reliability and credibility rating. Source reliability is conceptually linked to "confidence" in a given source, based on past performance, while information

[1] Funding support for this work was provided by the Canadian Safety and Security Program Project CSSP-2016-TI-2224 (Improving Intelligence Assessment Processes with Decision Science).

credibility reflects the extent to which new information conforms to previous reporting. It is also worth noting that NATO STANAG 2511 uses *confirmed by other sources* as its highest information credibility rating, where current Allied doctrine substitutes *completely credible.*

Table 7-1: NATO AJP 2.1 2016 Source Reliability and Information Credibility Scales [9].

Reliability of the Collection Capability		Credibility of the Information	
A	Completely reliable	**1**	Completely credible
B	Usually reliable	**2**	Probably true
C	Fairly reliable	**3**	Possibly true
D	Not usually reliable	**4**	Doubtful
E	Unreliable	**5**	Improbable
F	Reliability cannot be judged	**6**	Truth cannot be judged

Table 7-2: NATO STANAG 2511 Source Reliability Scale [10].

Reliability of Source		
A	Completely reliable	Refers to a tried and trusted source which can be depended upon with confidence.
B	Usually reliable	Refers to a source which has been successful in the past but for which there is still some element of doubt in a particular case.
C	Fairly reliable	Refers to a source which has occasionally been used in the past and upon which some degree of confidence can be based.
D	Not usually reliable	Refers to a source which has been used in the past but has proved more often than not unreliable.
E	Unreliable	Refers to a source which has been used in the past and has proved unworthy of any confidence.
F	Reliability cannot be judged	Refers to a source which has not been used in the past.

Critical examination of these standards and others collected by SAS-114 exposes a number of weaknesses and inconsistencies. Given the extensive influence of the Admiralty Code, and efforts by many Alliance members to conform to NATO doctrine, the issues outlined below are common across most of the standards examined.

7.2.1 Semantic Issues

Under the Admiralty Code, qualitative ratings of reliability and credibility form a demonstrably intuitive progression [11]. However, subjective interpretations of the boundaries between these ratings are likely to vary among users, as are interpretations of the relevant rating criteria [12]. For instance, in many versions of the Admiralty Code, a *reliable* ('A') source is said to have a "history of complete reliability," while a *usually reliable* ('B') source has a "history of valid information most of the time" [2], [13], [14], [15], [16], [17]. None of the standards examined associate these descriptions with numerical values (i.e., 'batting averages'),

potentially leading to miscommunication. One analyst may assign *usually reliable* to sources that provide valid information > 70% of the time. An analyst receiving this rating may interpret it to mean valid information > 90% of the time, and place more confidence in the source than is warranted. Conversely, an analyst may assume *usually reliable* reflects valid information only > 50% of the time, and prematurely discount the source. Asked to assign absolute probability values to reliability and credibility ratings, US intelligence officers demonstrated considerable variation in their interpretations [11]. For example, probabilistic interpretations of *usually reliable* and *probably true* ranged from .55 to .90 and .53 to .90, respectively, while interpretations of *fairly reliable* and *possibly true* both ranged from .40 to .80 [11].

Table 7-3: NATO STANAG 2511 Information Credibility Scale [10].

Credibility of Information		
1	Confirmed by other sources	If it can be stated with certainty that the reported information originates from another source than the already existing information on the same subject, it is classified as "confirmed by other sources" and is rated "1".
2	Probably true	If the independence of the source of any item or information cannot be guaranteed, but if, from the quantity and quality of previous reports its likelihood is nevertheless regarded as sufficiently established, then the information should be classified as "probably true" and given a rating of "2".
3	Possibly true	If, despite there being insufficient confirmation to establish any higher degree of likelihood, a freshly reported item of information does not conflict with the previously reported behaviour pattern of the target, the item may be classified as "possibly true" and given a rating of "3".
4	Doubtful	An item of information which tends to conflict with the previously reported or established behaviour pattern of an intelligence target should be classified as "doubtful" and given a rating of "4".
5	Improbable	An item of information which positively contradicts previously reported information or conflicts with the established behaviour pattern of an intelligence target in a marked degree should be classified as "improbable" and given a rating of "5".
6	Truth cannot be judged	Any freshly reported item of information which provides no basis for comparison with any known behaviour pattern of a target must be classified as "truth cannot be judged" and given a rating of "6". Such a rating should be given only when the accurate use of higher rating is impossible.

Among the standards examined, *reliable* or *completely reliable* indicates maximum source reliability, while *confirmed*, *confirmed by other sources*, or *completely credible* marks the highest degree of information credibility. Despite these inconsistencies, most scales faithfully reproduce the Admiralty Code's A – F (reliability) / 1 – 6 (credibility) scoring scheme, and ratings are often communicated using only the appropriate alphanumeric code (e.g., A1). These terminological variations may therefore contribute to miscommunication between users familiar with different standards. For instance, under most US standards examined [13], [14], [16], [17] 'A' is defined as *reliable*, while UK Joint Doctrine 2-00 [18] defines 'A' as *completely reliable* (conforming to NATO doctrine). A US analyst who understands 'A' to mean *reliable*

might transmit that rating to a UK counterpart, who interprets it as *completely reliable*. This translation is potentially problematic, given that an analyst or consumer may place more weight on a source labelled *completely reliable* than one labelled *reliable*. Alternatively, the translation from *completely reliable* to *reliable* could lead a recipient to undervalue a source.

Inter-standard miscommunication could also arise where scales use the term 'accuracy' as a synonym for information credibility (e.g., Refs. [14], [19]). While credibility often includes considerations of accuracy, it is likely a more multidimensional construct. Credibility generally incorporates criteria that can serve as cues to accuracy, but which are not equivalent to accuracy (e.g., triangulating evidence contributes to credibility, but does not require ground truth). Thus, this use of 'accuracy' by certain standards may further diversify interpretations of ratings, as well as the determinants considered during evaluation.

Another semantic issue relates to liberal use of terms conveying certainty (e.g., *confirmed*). In intelligence contexts where the information "is always incomplete... [and] frequently ambiguous," [20] these expressions could lead to overconfidence on the part of consumers. Compounding this issue is the observed tendency of analysts to confine their ratings to the high ends of the scales [21]. In their review of spot reports completed during a US Army field exercise, Baker, McKendry, and Mace [21] found that A1 and B2 represented 80% of all reliability/credibility ratings, with B2 alone comprising 74% of ratings. Allied intelligence doctrine explicitly discourages statements of certainty "given the nature of intelligence projecting forward in time" [9]. However, it remains unverified whether *completely credible* actually conveys less certainty than *confirmed by other sources*. A piece of information may be confirmed by some sources and simultaneously disconfirmed by others (an issue further explored in Section 7.2.3). Researchers could measure subjective interpretations of *completely credible* and compare them with interpretations of *confirmed by other sources*.

7.2.2 Source Reliability Determinants

To address miscommunication stemming from vague source history descriptors, this determinant could be quantified (e.g., source reliability = accurate information provided / total information provided). A quantitative method of tracking and updating source history could improve consistency and streamline the information evaluation process [22]. However, this would fail to address the Admiralty Code's implicit treatment of source reliability as constant across different contexts [12]. Regardless of past performance, source reliability may vary dramatically depending on the type of information provided, characteristics of the source(s), and the circumstances of collection. A Human Intelligence (HUMINT) source with a proven track record reporting on military operations may lack the expertise to reliably observe and report on economic developments. Beyond variable expertise, HUMINT source motivations, expectations, sensitivity, and recall ability may shift between situations, with major implications for information quality [23], [24]. Even the reliability of an 'objective source' (i.e., a sensor) is highly context dependent [25]. For example, inclement weather may compromise the quality of information provided by an optical sensor, despite a history of perfect reporting under ideal conditions. Future research could evaluate means of incorporating contextual information into ratings of source reliability, or into a more holistic measure of information quality.

Aside from source history, most of the standards examined highlight reliability determinants such as "authenticity," "competency," and "trustworthiness." The inclusion of these determinants is consistent with the broader literature on source reliability [23], [24]. However, the extant standards fail to formally define or operationalize these concepts. Their inclusion is therefore likely to increase subjectivity and further undermine the fidelity of reliability assessments. The standards examined also fail to operationalize the qualifiers used to describe each level. For instance, reliability ratings often incorporate whether an evaluator has "minor doubt," "doubt," or "significant doubt" about the source's authenticity. Aside from being vague, the use of modifiers ("minor," "significant") for some levels, and the unmodified term ("doubt") for another, is problematic because the unmodified term effectively subsumes the modified cases. Chang *et al.* [26] describe how a process

designed to decompose and evaluate components of a problem (i.e., information characteristics) may amplify unreliability in assessments if that process is ambiguous and open to subjective interpretations. Given the ambiguity built into current standards, users are unlikely to retrieve every relevant determinant, let alone reliably and validly weigh every relevant determinant when arriving at an ordinal assessment.

Another issue with current source reliability standards is their failure to delineate procedures for evaluating 'subjective sources' vs. 'objective sources' (e.g., human sources vs. sensors) [27], or primary sources vs. secondary/relaying sources [28]. A determinant such as source motivation may be relevant when assessing HUMINT sources, but not sensors. Similarly, source expertise may be highly relevant for a primary source collecting technical information (e.g., a HUMINT asset gathering information on Iranian nuclear technology), but less so for an intermediary delivering this information to a collector. In cases where information passes through multiple sources, there are often several intervals where source reliability considerations are relevant [28]. For instance, when receiving second-hand information from a HUMINT source, one might consider the reliability of the primary source, the reliability of the secondary/relaying source(s), the reliability of the collector, as well as the reliability of any medium(s) used to transmit the information [28].

Following initial collection, Nobel [29] describes how information may undergo distortion at other stages of the intelligence process. Just like sources, intelligence practitioners will vary in terms of their ability to reliably assess and relay information. For instance, an economic subject matter expert may lack the expertise to accurately evaluate and transmit information on enemy troop movements. Beyond expertise, an intelligence practitioner's assessment is also undoubtedly influenced by his/her personal characteristics (e.g., motivation, expectations, biases, recall ability) as well as various contextual factors [12], [23], [29]. When a finished intelligence product is edited and approved for dissemination, managers may inject additional distortion by adjusting analytic conclusions [29]. The many opportunities for distortion may warrant the formalization of information evaluation as an ongoing requirement throughout the intelligence process (see Section 7.4) [12]. At the very least, efforts should be made to ensure intelligence practitioners and consumers are cognizant of the mutability of information characteristics following the initial evaluation.

7.2.3 Information Credibility Determinants

Much like the source reliability standards examined, most of the information credibility scales suffer from an inherent lack of clarity. Information credibility generally incorporates confirmation "by other independent sources" as a key determinant. However, no guidance is provided as to how many independent sources must provide confirmation for information to be judged credible. Where one analyst considers confirmation by two sources sufficient for a *confirmed* rating, another might seek verification by three or more. Perceptions of how much corroboration is necessary may also vary depending on the information in question. For instance, an analyst may decide that a particularly consequential piece of information requires more corroboration than usual to be rated *confirmed*. This lack of consistency could lead analysts to misinterpret each other's credibility ratings, and consider pieces of information more or less credible than intended.

The information credibility standards examined also lack instructions for grading pieces of information that are simultaneously confirmed and disconfirmed. Under the Admiralty Code, such information could be considered both *confirmed / completely credible* ('1') and *improbable* ('5') [25]. Without guidance, some analysts may base their assessments more heavily on confirmed information, while others focus on disconfirmed information, or pursue a balance between confirmed and disconfirmed. These three approaches could generate very different assessments, despite evaluating the same information.

Capet and Revault d'Allonnes [12] argue that confirmation does not, in itself, translate into information credibility, and that not all forms of confirmation should be weighted equally. Theoretically, a spurious rumour corroborated by many unreliable sources (e.g., tweets about a second shooter during a terrorist attack), and disconfirmed by a single reliable source (e.g., a police statement indicating a single attacker),

could still be rated highly credible under current standards. Capet and Revault d'Allonnes [12] advocate identifying a threshold whereby information must be confirmed by a clear majority, and undermined by few or no sources, while accounting for source reliability. This would directly contravene the Admiralty Code's treatment of source reliability and information credibility as independent.

Lesot, Pichon, and Delavallade [30] note that current standards lack consideration of whether relationships of affinity, hostility, or independence[2] exist between corroborating sources. Corroboration from a source that has a 'friendly' relationship with the source under scrutiny should likely have less influence than corroboration from an independent or hostile source. For example, all else being equal, if Saudi Arabia corroborates information provided by Syria (with which it has a hostile relationship), that confirmation should carry more weight than identical confirmation provided by Russia (which has a relationship of affinity with Syria). As a general rule, friendly sources should be expected to corroborate each other [30].

Friedman and Zeckhauser [31] suggest that the current emphasis on consistency with existing evidence may encourage confirmation bias. "Biased attrition" is used to describe an information filtering process that systematically favours certain information types in a problematic way. Information that conflicts with prior beliefs and analysis may in fact be more valuable, as it can shift the views of analysts and consumers more significantly. Friedman and Zeckhauser [31] argue that credibility standards could reduce biased attrition by incorporating the extent to which information provides a new or original perspective on the intelligence requirement at hand. Capet and Revault d'Allonnes [12] also suggest that current standards be modified to gauge the extent to which information provides "meaningful" corroboration.

Along similar lines, Lemercier [28] notes that confirmation-based credibility standards do not account for the phenomenon of amplification, whereby analysts come to believe closely correlated sources are independently verifying a piece of information. In order to control for amplification, credibility evaluation could incorporate successive corroboration by the same source, corroboration by sources of the same type, as well as comparative corroboration from different collection disciplines [28].

The current emphasis placed on confirmation/consistency may also reinforce order effects, given that new information must conform to prior information to be deemed credible. All else being equal, if an analyst receives three new pieces of information, the first item received will typically face the fewest hurdles to being assessed as credible. Meanwhile, the second piece of information must conform to the first, and the third must conform to both the first and second. Under this system, an analyst may inadvertently underweight information that is in fact more accurate or consequential than information received earlier, potentially decreasing the quality of analysis. One option for dealing with order effects would be the formal inclusion of mechanisms to revaluate prior pieces of information as new information becomes available (a prospect that is further explored in Section 7.4). Two of the US standards examined [17], [19] advocate continuous analysis and re-evaluation of source reliability / information credibility as new information becomes available. However, neither document outlines a specific method for revaluation.

Beyond confirmation, most of the information credibility scales examined incorporate consideration of whether an item is "logical in itself." Current standards do not specify whether this simply refers to the extent that information conforms to the analyst's current assessment. Furthermore, the use of "not illogical" as a level between "logical in itself" and "illogical in itself" is nonsensical, as "not illogical" effectively means "logical" (in itself).

As noted with regards to source reliability, the Admiralty Code's one-size-fits-all approach to information credibility neglects important contextual considerations. Several US standards suggest that credibility determinants have more relevance depending on the collection discipline(s) utilized. For example,

[2] Several credibility standards do call for the independence of corroborating sources (e.g., Refs. [13], [15]), but none examined consider other types of relationships.

TC 2-91.8 [14] and ATP 2-22.9 [16] suggest that there is a greater risk of deception (an information credibility determinant) when utilizing Open-Source Intelligence (OSINT) than Captured Enemy Documents (CEDs). Similarly, ATTP 2-91.5 [19] refers to the Admiralty Code as the "HUMINT system," and recommends the development of separate rating systems to assess the three basic components of document and media exploitation (Document Exploitation [DOMEX], Media Exploitation [MEDEX], Cellphone Exploitation [CELLEX]).

Joseph and Corkill [32] stress that the Admiralty Code is a grading system rather than an evaluation methodology. Beyond what is outlined in the scales, evaluators may have a formal assessment procedure and/or a more exhaustive list of determinants to consider. Supplementary documents add some clarity to the standards examined, but also vary in terms of which determinants are identified and emphasized. Additionally, none of these extra determinants are defined or operationalized, and may further contribute to subjectivity. The following factors are highlighted in one or more of the documents examined:

Reliability:

- Circumstances under which information was obtained;
- Quality of source's bona fides; and
- Sensor capabilities.

Credibility:

- Internal and external consistency;
- Risk of denial and deception;
- Timeliness/recency; and
- Unusual absence of evidence.

Overlapping Factors[3]:

- Source access;
- Source expertise/authority; and
- Source motivation.

7.3 CONCEPTUALIZING INFORMATION QUALITY

As noted previously, the Admiralty Code is predicated on the independence of source reliability and information credibility. In developing a comprehensive, evidence-based means of information evaluation, an initial step would be to evaluate the independence of these constructs, and to assess whether they are unidimensional (e.g., credibility is understood as the probability that information is accurate) or multidimensional (e.g., credibility is understood as a profile of several determinants, including internal consistency, external consistency, timeliness, risk of deception, etc.) [12]. If the current terminology is found to be unidimensional, further experimentation could yield a list of qualitative terms with narrower and more consistent interpretations. Alternatively, if the meaning of the current terminology is multidimensional, this may warrant the creation of new scales to gauge factors comprising a comprehensive measure of information quality (e.g., information quality = a function of timeliness rated from 0 – 5; risk of deception rated from 0 – 5; source reporting history rated from 0 – 5, etc.). This latter approach would resemble the UK Defence Intelligence pilot approach to assessing analytic confidence.

[3] According to UK JDP 2-00 [18], these factors affect both source reliability and information credibility.

Several studies support the introduction of a single measure of perceived information quality (i.e., accuracy/truthfulness) incorporating all available information, including source reliability. Analysts are shown to pair reliability and credibility scores from the same level [8], [21] or to base decisions about accuracy more on credibility than reliability [22]. Nickerson and Feehrer [33] note that when no other information is available to gauge information credibility, analysts will logically base their rating on source reliability, given that reliable sources tend to produce credible information. To this point, Lemercier [28] posits that determining source reliability is not an end in itself, but rather a means of assessing information credibility, which he suggests is the ultimate goal of the evaluator.

A single measure of accuracy/truthfulness could address several challenges related to incongruent ratings and the lack of comparability between the two scales [12]. Samet [22] shows that analysts assign likely accuracy/truthfulness less reliably when basing their decision on separate reliability and credibility metrics, than a single measure. Similarly, in a preliminary analysis, Mandel, Dhami, Weaver, and Timms (cited in Ref. [34]) find that analysts show poor test-retest reliability when estimating the accuracy of information with incongruent reliability/credibility scores (e.g., A5, E1). Mandel *et al.* [34] also show that inter-analyst agreement plummets as source reliability and information credibility scores become less congruent. This suggests that while the two scales may be distinct in theory, in practice, users do not treat them as such. The ambiguity inherent in combining incongruent ratings may partially explain why evaluators often default to ratings from the same level [12].

Beyond issues stemming from incongruent ratings, currents standards also lack mechanisms for comparing multiple items of varying quality, which is a regular requirement for intelligence analysts [35]. For instance, it is unclear how analysts should weigh one piece of information rated B3 against another rated C2. The margin of interpretation may be increased by the use of two different scale types; credibility comprises a positive-negative scale (information is confirmed/invalidated), while reliability ranges from low/non-existent (the source has provided little/no credible information) to a maximum level (the source has a history of complete reliability) [12]. Without any sort of fusion methodology, this represents another sensitive process left to the subjective judgement of individual analysts [36]. The lack of comparability between scales also means that current measures of reliability and credibility are ill-suited for integration into an automated or semi-automated system for information evaluation [25]. A single, comprehensive measure of information quality would likely be more conducive to the collation of information of varying quality.

If quantified, such a measure could also enable users to grade information with finer discrimination. While Samet [11] finds that users can make quantitative distinctions between the five levels in each scale, the average size of the difference between the mean probabilities assigned by users to adjacent levels indicates there is room for greater precision. These findings are similar those of Friedman *et al.* [37] in the context of qualitative probability assessments. They find that analysts can assign probabilities more precisely than conventional wisdom supposes, and argue that the imprecision built into current standards sacrifices predictive accuracy. The inclusion of empirically grounded numerical values could also mitigate language barriers and some of the inter-standard semantic issues identified, while improving collaboration and analyst accountability [22].

In developing a comprehensive measure of information quality, it would be necessary to consider the hierarchy of relevant determinants, possible interactions or tradeoffs between determinants, as well as their importance depending on context and end-user information requirements [27], [38]. For example, confirmation may be less important than (or even conflict with) considerations of timeliness, where the pursuit of confirmation translates into unacceptable decision latency [38]. As noted previously, certain determinants (e.g., motivation) may be completely irrelevant depending on the information under scrutiny. In general, a determinant of information quality can be deemed relevant if a change of its value impacts the hypotheses under consideration; the levels of belief assigned to those hypotheses; or the utility values assigned to a set of potential courses of action [27].

Rogova [27] provides ontologies of quality of information content (Figure 7-1) and quality of information sources (Figure 7-2) that were derived from the broader literature on information fusion for decision support. Detailed explanations of each concept within these ontologies and the evidential basis for the concepts are beyond the scope of this chapter but can be found in Ref. [27]. Our purpose in presenting these ontologies is simply to highlight that a comprehensive measure of information quality developed for intelligence analysis might incorporate many of the determinants identified. These models also depict the interconnectedness of information characteristics, which current standards fail to address. Researchers formulating an intelligence-oriented model of information quality could evaluate the prospects of combining these ontologies, and the utility of other quality measures identified in the information fusion scholarship.

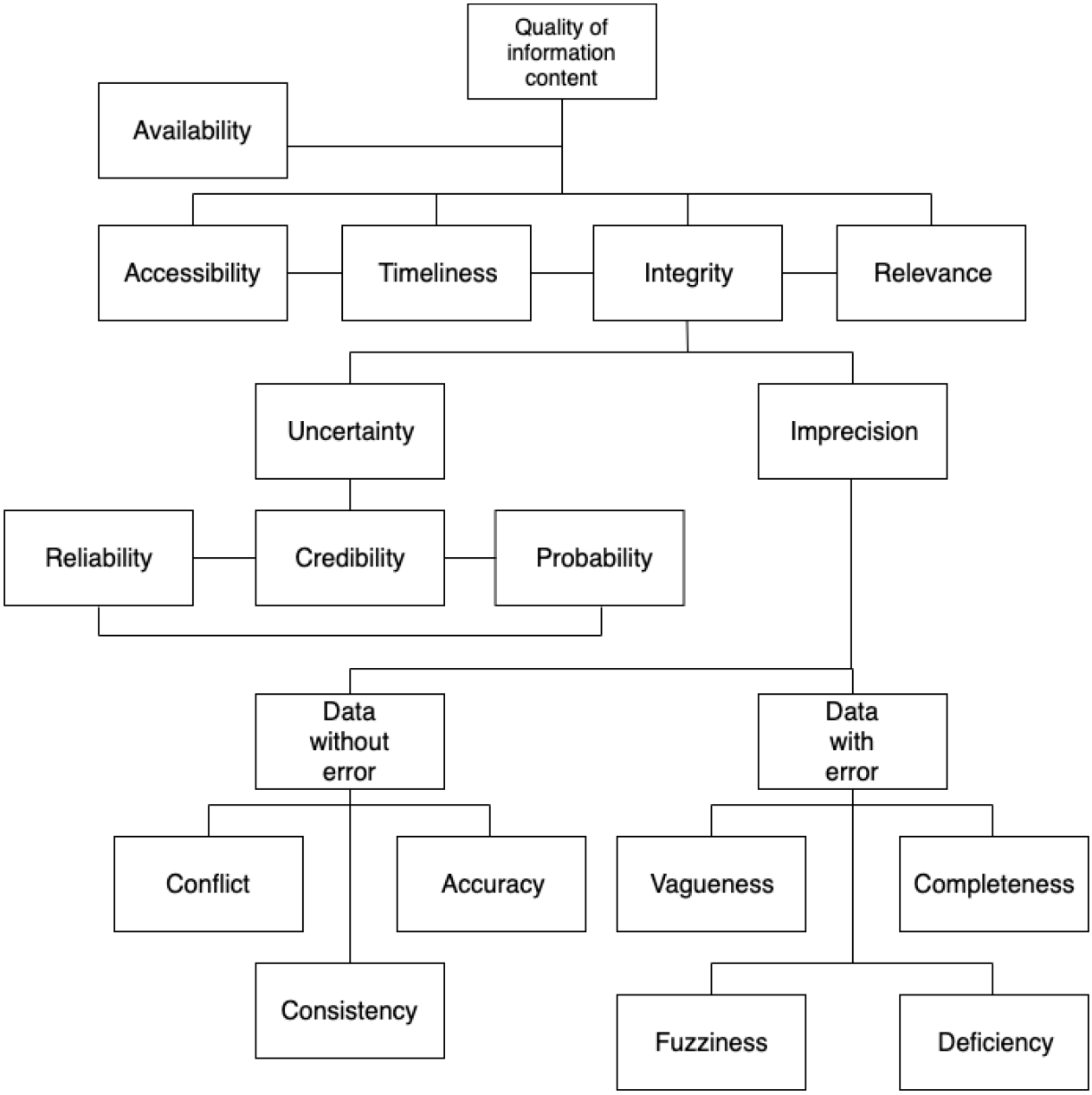


Figure 7-1: Rogova's Ontology of Quality of Information Content [27].

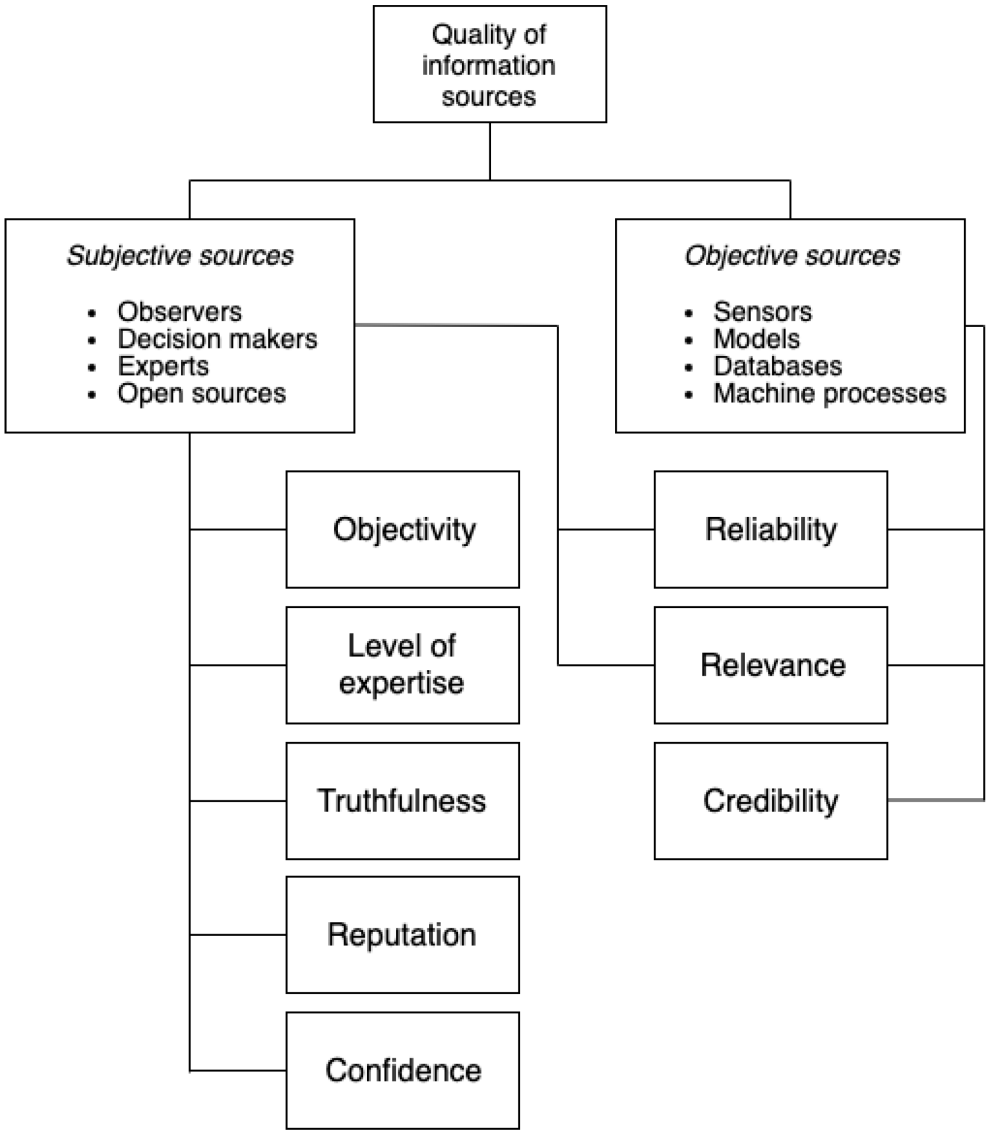


Figure 7-2: Rogova's Ontology of Quality of Information Sources [27].

7.4 ALTERNATIVE APPROACHES TO INFORMATION EVALUATION

After determining the relationship between source reliability and information credibility, and formulating a comprehensive measure (or measures) of information quality, the next task for researchers would be to identify possible models for evaluation. A wide range of information evaluation systems have been proposed, but few have undergone rigorous empirical evaluation, particularly in intelligence contexts [4], [35], [39], [40]. Once promising models are identified, researchers could evaluate them using intelligence practitioners and realistic intelligence problems.

Before reviewing alternative approaches to information evaluation, it is worth noting the variation among current standards in terms where information evaluation is situated within the intelligence process. For instance, NATO intelligence doctrine [9] embeds evaluation procedures within the processing stage, thus emphasizing the analyst's role in gauging information characteristics. UK JDP 2-00 [18] outlines a joint role for analysts and collectors, whereby collectors pre-rate information characteristics before analysts weigh in with their own (potentially broader) understanding of the subject. Refs. [3], [24], and [29] stress the primary

collector's role in assessing reliability, particularly in contexts where access to a clandestine source is restricted to the agent handler. This inconsistency is significant given the noted mutability of information characteristics over time, across contexts, and at different stages of the intelligence process itself [23], [24], [25], [29]. Whether information is assessed upon initial collection, by an analyst during processing, or by several practitioners throughout the intelligence process could have a substantial impact on its reliability/credibility evaluation. Consequently, the extent to which information is deemed fit to use will largely determine its influence (or lack thereof) on analytic judgements. In other words, the timing of source and information evaluation within the intelligence process could add additional inter-analytic unreliability to such meta-informational ratings and, by proxy, to analytic judgements themselves. For instance, an intriguing question is whether information evaluation is given more weight in intelligence analysis when the evaluation step is also conducted by the analyst rather than by the collector. Moreover, do individual differences in analyst characteristics play a role in how the source of meta-information is treated in subsequent analysis. Perhaps analysts who have a disposition of high self-confidence place more weight in such evaluations when they rendered them, whereas analysts who are perennial self-doubters might give more weight to evaluations that come from other sources. These hypotheses could be tractably tested in future research.

A compounding issue is the absence of mechanisms for revaluation when new information becomes available and determinants, such as a source's reliability rating, are updated [28]. For example, under current standards, it is unclear how users should treat information provided by a source long considered *completely reliable*, but suddenly discovered to be *unreliable*. This is particularly complicated when information ratings form an interdependent chain (e.g., Info A's rating is tied to Info B's rating; Info B's rating is tied to the rating of Source X; Source X has just been exposed as a double agent). Together, these issues may warrant the implementation of information evaluation as an iterative function throughout the intelligence process. This approach could be applied to the evaluation of individual pieces of information [4], as well as the marshalling of evidence when forming analytic judgements [35]. As noted, certain US standards [17], [19] advocate continuous revaluation of information quality, but none of them provide formalized mechanisms for doing so.

Bayesian networks could represent one means of capturing complicated interactions between items of evidence, which may influence each other's reliability/credibility [1]. As new information becomes available, Bayesian networks can be updated coherently; that is, respecting the axioms of probability theory, such as unitarity, additivity, and non-negativity [41]. This process may reduce the systematic errors exhibited by individuals estimating the impact of less than totally reliable information [1]. Integrating Bayesian methods (or probabilistic approaches, more generally) into information evaluation could also improve the fidelity of assessments where sources communicate information using probabilistic language. Current standards provide no guidance for incorporating probabilistic expressions into a broader evaluation of information quality determinants (e.g., if a *usually reliable* source reports that she *probably* saw two helicopters). Given the subjectivity inherent in interpreting probability phrases commonly used in intelligence production (for discussion, see Ref. [42]), this could be another source of miscommunication embedded in current standards.

McNaught and Sutovsky [35] propose using a Bayesian network as a computer-assisted framework to facilitate evidence marshalling, and the fusion of information of varying quality. While they suggest that such networks may help analysts explore uncertain situations and overcome cognitive biases, they warn that routine (as opposed to supplemental) use of these models could generate inaccuracy due to the challenges of estimating certain input parameters (e.g., quantifying the reliability of a HUMINT source, especially under conditions of anonymity). To this point, Rogova [27] argues that *a priori* domain knowledge is often essential when determining many of the input parameters in a system for assessing information quality. McNaught and Sutovsky [35] only advocate the use of Bayesian networks for evidence marshalling where the input parameters are known with a "reasonable degree" of accuracy. Simply put, a coherent integration of "garbage in", which Bayesian approaches should ensure, will still yield "garbage out."

The complete automation of information evaluation may be undesirable, given the requirement for analysts to easily understand and modify their inputs as new information becomes available [30]. To this end, Lesot, Pichon, and Delavallade [30] propose a semi-automated method for evaluating information derived from textual documents, which is based on a possibility framework for managing uncertainty resembling the structured analytic technique known as Analysis of Competing Hypotheses (ACH) [43]. Their method first identifies pieces of information relevant to the requirement at hand, and then attaches an independent level of confidence to each piece of information. These ratings are then combined to calculate an overall degree of confidence, which the analyst can attach to judgements derived from all available information. Lesot, Pichon, and Delavallade [30] suggest that their method automates a large portion of information evaluation, enabling the processing of large volumes of data, while giving analysts control over each stage of the process. Simulating a situation where high-quality information is provided by relatively reliable sources, they demonstrate the ability of the proposed method to identify an optimal aggregation operator with an information fusion function. They note that this process requires further evaluation under different conditions, as well as a review by domain specialists.

An alternative model is the computational Method for Assessing the Credibility of Evidence (MACE), designed by Schum and Morris [4] for application in HUMINT contexts. Incorporating both Baconian and Bayesian methods, MACE draws on procedures from the Anglo-American legal tradition for gauging the competence and credibility of witnesses. MACE first guides users through a Baconian analysis to assess how much evidence is available about a particular source, and how completely source competence and credibility (i.e., reliability) can be evaluated. Users answer 25 sequential questions related to source competence, veracity (the extent to which a source believes the information being relayed is true), objectivity, and observational sensitivity. The system tracks inputs for each question, as well as any questions that remain unanswered. While subjective, the final output (which also resembles an ACH matrix), is evidence-based and can be easily updated as more information becomes available. Referencing these scores, the second stage of MACE guides users as they generate three pairs of likelihood estimates related to source characteristics, which are plotted on a two-dimensional probability space. MACE then applies Bayesian probability methods to estimate the strength of evidence about a particular HUMINT source. The process yields an assessment of posterior odds favouring the believability of the source. In this second stage, MACE again enables users to easily update and modify their assessments. While MACE focuses on HUMINT sources, the question set could be extended to evaluate other types of information [44].

The use of such approaches could also be supplemented with training designed to improve collectors' and analysts' applied statistical skill. For example, Mandel [45] designed a brief (approximately 30-minute) training protocol on Bayesian belief revision and hypothesis testing with probabilistic information. Intelligence analysts were assessed on the accuracy and probabilistic coherence before and after receiving the training. Mandel [45] found statistically significant improvement after training on both accuracy and coherence of analysts' probability estimates, suggesting that intelligence professionals can reap quick wins in learning that might enable them to better understand the kinds of probabilistic models noted in the aforementioned examples. Similar encouraging results have been reported elsewhere [46], [47]. This type of training is not only important for understanding such models, however. People routinely violate logical constraints on probability assessments (e.g., see Refs. [41] and [48], [49], [50], [51]), and there is no good reason to believe that analysts are exempt. Indeed, the findings of Ref. [45] show that they are not.

The aforementioned evaluation systems are by no means exhaustive. Other mechanisms for information evaluation and the challenges they present can be found in Refs. [35], [39], [40]. Each system requires further experimentation. Furthermore, each system responds to a different analytic challenge: either the need to integrate information quality into analytic judgements [35]; to collate information of varying quality [30]; or to evaluate the quality of information *about* a source, as well as the quality of the source itself [4]. Despite these differences, each system is iterative, easily updated, and designed to be more reliable, transparent, and comprehensive than current evaluation standards. Researchers and intelligence practitioners ought to examine these and other systems as they pursue more effective methods of information evaluation in

intelligence production. More generally, the intelligence community would be well served by taking a more evidence-based approach to verifying the effectiveness of its current methods and improving upon those where possible [52], [53], [54]. For far too long it has relied on developing analytic tradecraft methods that merely have to pass the test of apparent plausibility and effectiveness. Going forward, it should proactively leverage relevant information – theory, findings, and methods – from judgement and decision science.

7.5 REFERENCES

[1] Johnson, E.M., Cavanagh, R.C., Spooner, R.L., and Samet, M.G. (1973). Utilization of reliability measurements in Bayesian inference: Models and human performance. *IEEE Transactions on Reliability* 22 (3):176-182.

[2] United Nations Office on Drugs and Crime. (2011). *Criminal Intelligence Manual for Analysts*. Vienna, Austria.

[3] Carter, D.L. (2009). *Law Enforcement Intelligence: A Guide for State, Local, and Tribal Law Enforcement Agencies*, (2nd ed.). Washington DC: Office of Community Oriented Policing Services, US Department of Justice.

[4] Schum, D.A., and Morris, J.R. (2007). Assessing the competence and credibility of human sources of intelligence evidence: Contributions from law and probability. *Law, Probability and Risk,* 6(1-4):247-274.

[5] Betts, R.K. (2008). Two faces of intelligence failure: September 11 and Iraq's missing WMD. *Political Science Quarterly* 122 (4):585-606.

[6] Hanson, J.M. (2015). The admiralty code: A cognitive tool for self-directed learning. *International Journal of Learning, Teaching and Educational Research* 14 (1):97-115.

[7] United States Department of the Army. (1951). *Field Manual FM 30-5, Combat Intelligence*. Washington DC.

[8] Miron, M.S., Patten, S.M., and Halpin, S.M. (1978). *The Structure of Combat Intelligence Ratings*. Technical Paper 286. Arlington, VA: US Army Research Institute for Behavioral and Social Sciences.

[9] North Atlantic Treaty Organization (2016). *Allied Joint Doctrine for Intelligence Procedures AJP-2.1*. Brussels, Belgium.

[10] NATO Standardization Office. (2003). *STANAG 2511 – Intelligence Reports*, (1st ed.) Brussels, Belgium.

[11] Samet, M.G. (1975). *Subjective Interpretation of Reliability and Accuracy Sales for Evaluating Military Intelligence*. Technical Paper 260. Arlington, VA: US Army Research Institute for Behavioral and Social Sciences.

[12] Capet, P., and Revault d'Allonnes, A. (2014). Information evaluation in the military domain: Doctrines, practices, and shortcomings. In: *Information Evaluation*, Capet, P., and Delavallade, T. (Eds.), 103-125. Hoboken, NJ: Wiley-ISTE.

[13] United States Department of the Army. (2006). *Field Manual FM 2-22.3, Human Intelligence Collector Operations*. Washington DC.

[14] United States Department of the Army. (2010). *Training Circular TC 2-91.8, Document and Media Exploitation*. Washington DC.

[15] Department of National Defence. (2011). *Canadian Forces Joint Publication CFJP 2-0, Intelligence*. Ottawa, ON.

[16] United States Department of the Army. (2012). *Army Techniques Publication ATP 2-22.9, Open-Source Intelligence*. Washington DC.

[17] United States Department of the Army. (2012). *Army Techniques Publication ATP 3-39.20 Police Intelligence Operations*. Washington DC.

[18] United Kingdom Ministry of Defence. (2011). *Joint Doctrine Publication JDP 2-00, Understanding and Intelligence Support to Joint Operations*, (3rd ed.) Swindon, UK. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/311572/20110830_jdp2_00_ed3_with_change1.pdf.

[19] United States Department of the Army. (2010). *Document and Media Exploitation Tactics, Techniques, and Procedures ATTP 2-91.5* – Final Draft. Washington DC.

[20] Tecuci, G., Boicu, M., Schum, D., and Marcur, D. (2010). *Coping with the Complexity of Intelligence Analysis: Cognitive Assistants for Evidence-Based Reasoning*. Research Report #7, Learning Agents Center. Fairfax, VA: George Mason University.

[21] Baker, J.D., McKendry, J.M., and Mace, D.J. (1968). *Certitude Judgements in an Operational Environment.* Technical Research Note 200. Arlington, VA: US Army Research Institute for Behavioral and Social Sciences.

[22] Samet, M.G. (1975). Quantitative interpretation of two qualitative scales used to rate military intelligence. *Human Factors* 17 (2):192-202.

[23] Schum, D.A. (1987). *Evidence and Inference for the Intelligence Analyst*. Lanham, MD: University Press of America.

[24] Pechan, B.L. (1995). The collector's role in evaluation. In: *Inside CIA's Private World: Declassified Articles from the Agency's Internal Journal*, Westerfield, H.B. (Ed.), 99-107. New Haven, CT: Yale University Press.

[25] Cholvy, L., and Nimier, V. (2003). Information evaluation: Discussion about STANAG 2022 recommendations. In: *Proceedings of the NATO-IST Symposium on Military Data and Information Fusion.* Prague, Czech Republic.

[26] Chang, W., Berdini, E., Mandel, D.R. and Tetlock, P.E. (2018). Restructuring structured analytic techniques in intelligence. *Intelligence and National Security* 33 (3):337-356.

[27] Rogova, G.L. (2016). Information quality in information fusion and decision making with applications to crisis management. In: *Fusion Methodologies in Crisis Management, Higher Level Fusion and Decision Making*, Rogova, G.L., and Scott, P. (Eds.), 65-86. Cham, Switzerland: Springer International Publishing.

[28] Lemercier, P. (2014). The fundamentals of intelligence. In: *Information Evaluation*, Capet, P., and Delavallade, T. (Eds.), 55-100. Hoboken, NJ: Wiley-ISTE.

[29] Noble, G.P., Jr. (2009). Diagnosing distortion in source reporting: Lessons for HUMINT reliability from other fields. Master's thesis. Erie, PA: Mercyhurst College.

[30] Lesot, M., Pichon, F., and Delavallade, T. (2014). Quantitative information evaluation: Modeling and experimental evaluation. In: *Information Evaluation*, Capet, P., and Delavallade, T. (Eds.), 187-228. Hoboken, NJ: Wiley-ISTE.

[31] Friedman, J.A., and Zeckhauser, R. (2012). Assessing uncertainty in intelligence. *Intelligence and National Security* 27 (6):824-847.

[32] Joseph, J., and Corkill, J. (2011). Information evaluation: How one group of intelligence analysts go about the task. In: *Fourth Australian Security and Intelligence Conference*. Perth, Australia.

[33] Nickerson, R.S., and Feehrer, C.E. (1975). *Decision Making and Training: A review of Theoretical and Empirical Studies of Decision Making and Their Implications for the Training of Decision Makers*. Technical Report NAVTRAEQUIPCEN 73-C-0128-1. Cambridge, MA: Bolt, Beranek and Newman, Inc.

[34] Mandel, D.R. (2018). *Proceedings of SAS-114 Workshop on Communicating Uncertainty, Assessing Information Quality and Risk, and Using Structured Techniques in Intelligence Analysis*. NATO Meeting Proceedings. Brussels, Belgium: NATO STO.

[35] McNaught, K. and Sutovsky, P. (2012). Representing variable source credibility in intelligence analysis with Bayesian networks. In: *Fifth Australian Security and Intelligence Conference*, 44-51. Perth, Australia.

[36] Cholvy, L. (2004). Information evaluation in fusion: A case study. In: *Proceedings of the International Conference on Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2004*. Perugia, Italy.

[37] Friedman, J.A., Baker, J.D., Mellers, B.A., Tetlock, P.E., and Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly* 62 (2):410-422.

[38] Rogova, G., Hadzagic, M., St-Hillaire, M., Florea, M., and Valin, P. (2013). Context-based information quality for sequential decision making. In: *Proceedings of the 2013 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. San Diego, CA.

[39] Capet, P., and Delavallade, T. (2014). *Information Evaluation*. Hoboken, NJ: Wiley-ISTE.

[40] Rogova, G., and Scott, P. (2016). *Fusion Methodologies in Crisis Management Higher Level Fusion and Decision Making*. Cham, Switzerland: Springer International Publishing.

[41] Karvetski, C.W., Olson, K.C., Mandel, D.R., and Twardy, C.R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis,* 10 (4):305-326.

[42] Irwin, D., and Mandel, D.R. (2018). *Methods for Communicating Estimative Probability in Intelligence to Decision-Makers: An Annotated Collection*. DRDC Scientific Letter DRDC-RDDC-2018-L017. Toronto, ON: DRDC.

[43] Heuer, R.J., Jr. (1999). The *Psychology of Intelligence Analysis*. Washington DC: Central Intelligence Agency, Center for the Study of Intelligence.

[44] Tecuci, G., Schum, D.A., Marcu, D., and Boicu, M. (2016). *Intelligence Analysis as Discovery of Evidence, Hypotheses, and Arguments: Connecting the Dots*. New York, NY: Cambridge University Press.

[45] Mandel, D.R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Frontiers in Psychology*, 6:387.

[46] Sedlmeier, P., and Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General* ,130 (3):380-400.

[47] Chang, W., Chen, E., Mellers, B., and Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making,* 11(5):509-526.

[48] Tversky, A., and Koehler, D.J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review,* 101(4):547-567.

[49] Villejoubert, G., and Mandel, D.R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory & Cognition*,30(2):171-178.

[50] Mandel, D.R. (2005). Are risk assessments of a terrorist attack coherent? *Journal of Experimental Psychology: Applied*, 11(4):277-288.

[51] Mandel, D.R. (2008). Violations of coherence in subjective probability: A representational and assessment processes account. *Cognition*, 106(1):130-156.

[52] Pool, R. (2010). *Field Evaluation in the Intelligence and Counterintelligence Context: Workshop Summary*. Washington DC: The National Academies Press.

[53] Dhami, M.K., Mandel, D.R., Mellers, B.A., and Tetlock, P.E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science,* 10(6):753-757.

[54] Mandel, D.R. (2019). Can decision science improve intelligence analysis? In: *Researching National Security Intelligence: National Security Intelligence: Multidisciplinary Approaches.* Coulthart, S., Landon-Murray, M., and Van Puyvelde, D. (Eds.), 117-140. Washington, DC: Georgetown University Press.

Chapter 8 – A RELIABILITY GAME FOR SOURCE FACTORS AND SITUATIONAL AWARENESS EXPERIMENTATION[1]

Francesca de Rosa and Anne-Laure Jousselme
NATO STO Centre for Maritime Research and Experimentation (CMRE)
ITALY

Alessandro De Gloria
University of Genoa
ITALY

8.1 INTRODUCTION

No unique definition for the term "serious game" exists, however it appears that there is a strong agreement on the fact that serious games are "designed for a primary purpose other than pure entertainment" [1]. Most of the research and applications in the domain of serious games have mainly focused on education, training and user learning objectives (e.g., Refs. [2], [3], [4]). Different fields of engineering have started to look at serious games, not only from an educational perspective, but also as supporting design tools. A recent literature review of games used in engineering research [5] presents a classification of such games following the Gameplay/Purpose/Scope (G/P/S) taxonomy [6]. This model proposes game play, purpose and scope as relevant classification dimensions. The first dimension differentiates between play-based or game-based games. The former are games characterised by a lack of well-defined objectives and rules. The latter instead, have defined objectives and rules. The purpose dimension allows classifying games on the basis of their function (e.g., message broadcasting, training, data exchange). Message broadcasting games are the ones that have been developed with the aim of broadcasting a message (e.g., educative games, informative games, persuasive games and subjective games). Games for training are developed with the purpose of improving players (cognitive or physical) performances. Finally, data exchange games have the specific purpose of supporting data exchange, such as "collecting information from […] players" [5]. The scope dimension refers to the game market (e.g., state and government, military and defence, healthcare, education, corporate, religious, culture and art, ecology, politics, humanitarian, advertising, scientific research) and the target audience (e.g., professionals or general public).

From the above-mentioned review, it appears that in the realm of engineering research serious games focus has shifted from message broadcasting and training to the data exchange purpose. Two notable examples explore human problem-solving strategies to support computational algorithm optimisation in the context of protein structure design [7] and vehicle powertrain controller design [8] respectively. The findings derived from the use of the two games have shown that human-derived strategies can be a valuable resource when used in conjunction with computational algorithms.

Situational Awareness (SAW) is defined [9] as "the perception of the elements in the environment within a volume of time and space, comprehension of their meaning and the projection of their status in the near future" and the cognitive process that enables Situational Awareness as Situational Assessment (SA). Some serious games in the context of SAW have been developed (e.g., Ref. [10]). However, to the best of the authors' knowledge, the focus of such games remains on training and message broadcasting, with the exception of Ref. [11], which presents a game for assessing team SAW.

With the final goal of informing the design of multi-source information fusion systems, we present in this chapter a data exchange game, characterised by a game-based approach. The scope of this game, called the Reliability Game, is scientific research [6] as it aims at collecting data to be used in further research of

[1] This chapter is reprinted with permission from de Rosa et al. [12].

source factors impact on human SA and consequent SAW. The term *source factor* is used in this chapter with the specific meaning of element that characterises a source of information, such as its type (e.g., radar, human operator and historical databases), quality, reliability or attractiveness. It should be noted that although the target audience is professionals [6], namely Subject Matter Experts (SMEs) in Maritime SAW, it could be extended to general public through the development of an appropriate scenario.

In this chapter, after providing some details with respect to the motivation of the game (Section 8.2) and the notion of source reliability (Section 8.3), we will discuss the design approach and choices (Section 8.4). We will present the game outcomes, which demonstrate the effectiveness of the game mechanics (Section 8.5). Finally, the conclusions and the way ahead are discussed (Section 8.6).

8.2 MOTIVATION

Decision support tools take greater and greater importance in daily life, whether it maybe a common car navigator or more complex surveillance systems that support operators in different working environments such as safety, security, crises management, health and first aid. Those systems all aim at improving the user's SAW by helping the user to better capture, understand and predict future states of the situation at hand, for more informed decisions. SAW and the corresponding information systems form an important building block of the dynamic decision-making processes [13]. Although those systems provide support to humans in problem solving and decision making, they still remain an "enabler, facilitator, accelerator and magnifier of human capability, not its replacement" [14].

To obtain from data, information and finally insight, processing capabilities have to rely on human assets to work correctly. Indeed, the information value to a specific application cannot be decoupled from the human component, in terms of the ability to search, analyse and interpret the data or information provided. For example, it has been demonstrated that SA might be negatively impacted by system automation that drives operators out of the loop [15]. Moreover, within operational environments human can concurrently and interchangeably act in more than one of the following roles: "decision maker, monitor, information processor, information encoder and storer, discriminator, pattern recognizer [,] . . . ingenious problem solver" [16] or disseminator. To improve human-machine synergy and user acceptance the system underlying reasoning and communication schemes should be intelligible [17] and possibly intuitive to the human.

In addition to the challenge of a desirable human-centred system design approach [18], the systems that support SAW have to deal with an increasing volume and velocity of the information, coupled with an increase in the variety of the information and corresponding sources with a potential lack of veracity. Data and information fusion technologies come into play to support operators' SA and reduce the information overload. To this end, information aggregation (e.g., data fusion) approaches have proven to be effective, provided that the outputs are presented in an intuitive and actionable format that engenders trust [19], [20].

To get full advantage of the variety of sources beyond the ones traditionally in use, we need not only to combine them but also to correctly account for source factors in fusion processes [21], [22], [23]. With respect to source reliability, most mathematical fusion operators assume that the sources are fully reliable or at least equally reliable and therefore assign an equal weight on the resulting combined belief assessment [24]. In reality this assumption is not always satisfied and sources can differ in reliability. Several strategies within different uncertainty frameworks (e.g., Bayesian, belief functions) have been proposed to account for partially reliable sources [24]. Generally, the consideration of source reliability in the fusion process relies on discounting, pruning or reinforcement operations [21], [22], [23], [24], allowing, for instance, to completely discard a piece of information provided by an unreliable source or to strengthen the weight of information originating from a highly reliable source. However, further research is needed to clarify some concepts related to source reliability, to clarify the semantics of source quality

dimensions and to ensure that the implementation of those reliability accounting strategies in current support systems meet some criteria of understandability or intuition.

A literary review on human factors methods [25] lists several methods for the assessment of SAW. However, none of those techniques is suitable for our work, as they do not specifically focus on the SA, which instead is at the core of the Reliability Game [26]. In fact, the purpose of this innovative approach in the context of SAW experimentation is to collect data regarding players' belief changes as a function source factors; more specifically, source type and quality. To gather this data each player is presented with a scenario and plays several rounds of the game, where the only variation consists in the knowledge regarding source type and quality. The corresponding belief changes are captured though game items' position (cards) and final confidence assessment, as it will be explained in further details in the next sections.

8.3 THE RELIABILITY CONCEPT

For a proper consideration of reliability in the fusion process, it is helpful to understand what source reliability is and to define the underpinning elements that are central to its quantification. There is no universal definition of source reliability and even fields that have traditionally been working with multi-source information, such as military intelligence, have come neither to a definition, nor a formalisation of the concept, nor to an agreement on the rating of the source reliability [27]. Following Ref. [28], reliability is defined as "ability to rely on or depend on, as for accuracy, honesty and achievements". It is important to underline that the term ability does not represent an ability of the source itself, but rather our bet on the ability to rely on it. Therefore, it is our own estimate, which is a function of many factors including the capacity and/or willingness of the source of providing good information. In the field of intelligence, source reliability is evaluated on the basis of past meta-knowledge and experience with the specific source. However, in general, it might depend on several other factors, such as similarity, perceived expertise, attractiveness [29] of the source or experience with analogous sources (encapsulated in source type). The purpose of work described in the following sections is to understand how source factors underpinning source reliability – specifically, source type and source quality, impact SA and SAW. The source reliability is treated as a latent variable. Therefore, is never specifically mentioned in the Reliability Game execution.

8.4 THE RELIABILITY GAME OVERALL DESIGN

The Reliability Game core is reasoning under uncertainty with information provided by sources of different type and quality, which are assumed as two underpinning factors of source reliability. The aim of the Reliability Game is to capture the impact of source factors on human SA. One of the final goals is to inform the design of automated reasoners to be included in multi-source information fusion systems.

This section summarises the design of the Reliability Game. More specifically, the following subsections provide details about the world design (Section 8.4.1), the system design (Section 8.4.2) and the content design (Section 8.4.3). The world design is defined in Ref. [30] as "the creation of the overall backstory, setting and theme", while the "creation of rules and underlying mathematical patterns" is identified under the definition of system design [31]. Finally, with the term content design, we refer to the "creation of characters, items, puzzles and missions" [30].

The Reliability Game design follows a mechanic-driven approach, which starts from the definition of the game core, followed by the selection of specific Game Mechanics (GMs). The following GMs have been identified in an early stage of development:

[GM1] Assessment of hypotheses relative to a missing vessel; and

[GM1] Use of cards to communicate messages to the player.

Those two mechanics were selected as they proved to be effective elements proposed in the Risk Game [31] which is a game that aims at eliciting experts' knowledge regarding reasoning about concurrent events when dealing with information that differs in nature (e.g., from sensors or from humans) and in quality. The information quality dimensions considered in the Risk Game were accuracy, precision and trueness. Following Ref. [32], the term accuracy can be interpreted as the "closeness of agreement between a test result or measurement result and the true value". The term precision refers to the "closeness of agreement between independent test/measurement results obtained under stipulated conditions" and the trueness refers to "closeness of agreement between the expectation of a test result or a measurement result and a true value", where the measurement is the information. Unlike the Risk Game, the Reliability Game assumes as fixed such dimensions in order to obtain a more rigid experiment control. This choice has been driven by the need to isolate the experimental variables' (i.e., source type and source quality) impact on the players' belief assessment.

8.4.1 World Design

The game is set in a maritime scenario and refers to a fictitious geographical area partitioned into three sovereign countries:

[L1] Right Land is a failed and poor state;

[L2] Centre Land suffers from disorders due to the vicinity to Right Land; and

[L3] Left Land is a stable and rich country, thanks to the presence of oil extraction facilities within their Exclusive Economic Zone (EEZ) and to the Left Land Canal, which is a strategic waterway owned, managed and maintained by the Left Land government.

The player is part of Left Land Maritime Authority, which is the only authority with responsibilities within Left Land national waters. Therefore, it is responsible for maritime safety, maritime security, environmental protection, customs and port state control. More specifically the player embodies the head of the monitoring department, who is informed by a subordinate that the Automatic Identification System (AIS) contact of the tanker ship MV Red Horizon (Figure 8-1) has been lost since six hours. The player is asked to assess what is currently happening to the ship in order to take further actions.

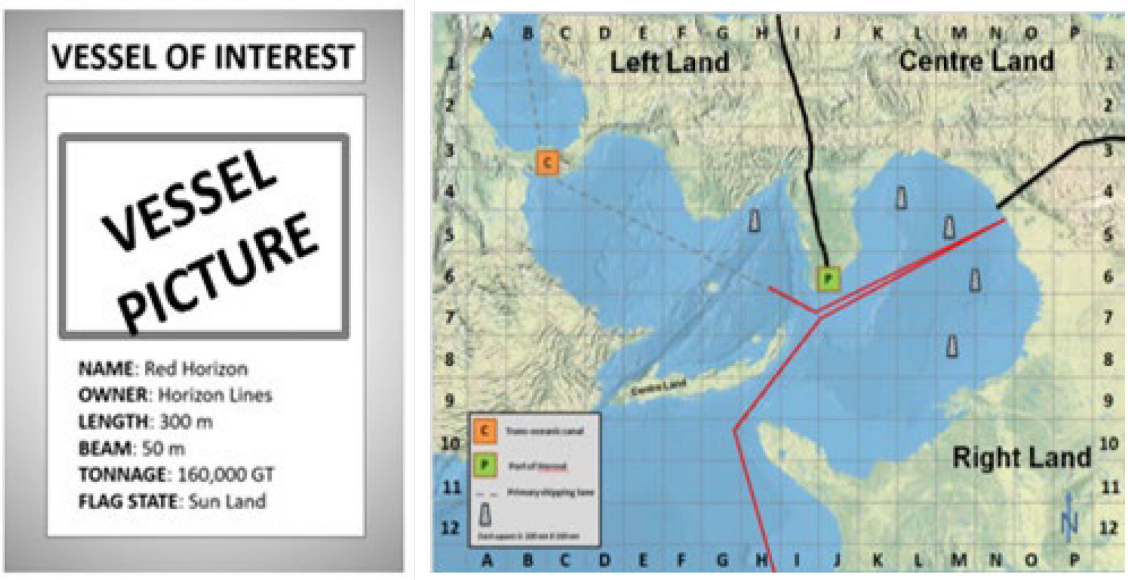


Figure 8-1: MV Red Horizon Information (Vessel of Interest) and Its Track Before AIS Contact Loss as Displayed in the Scenario Map by the Red Line.

The player is presented with a set of three mutually exclusive and collectively exhaustive hypotheses and is asked to perform a belief assessment about what is happening to the ship on the basis of the incoming information. The three candidate hypotheses for this scenario are:

[H1] Nothing happened: the ship is continuing its voyage without safety or security issues;

[H2] Safety issue: there is a safety issue with the ship; and

[H3] Security issue: the ship is engaged in oil smuggling activities.

Hypothesis H1 would be explained by the fact that the AIS signal is not received due to a possible failure of the AIS, and no intervention would be required. On the contrary, H2 or H3 would trigger respectively a Search and Rescue (SAR) operation or a security operation. The action phase *per se* is not part of the game as the game stops after the SA phase.

8.4.2 System Design

A game session is divided in four rounds, in which a set of eleven cards is provided to the player. In each round the player is requested to assess what is happening to the ship on the basis of the available information and meta-information on source factors (source type and source quality) provided through cards. We will refer to the card as conveying a Message (*M*), which is composed by the information *(I)* and associated meta-information about source factors, namely source quality (*Q*) and (*T*). The information provided might be true or false. Although it is not explicitly requested to assess information trueness, the player will implicitly assess this information dimension as a consequence of the game dynamics.

A summary of the game state, intended as the picture of all relevant variables that may change during the play [30] is reported in Table 8-1. Table 8-2 summarises the game view, which is the portion of the game state that is visible to the player in each round [30]. With respect to the game view it can be noticed that each round is exactly the same (e.g., scenario, triggering event, information presented in the cards, order of cards), with the only exception of the meta-information about source factors (*Q* and *T*). The order of the cards is kept constant with the purpose of controlling the information presentation order effect [33].

Table 8-1: Reliability Game State.

Variable	Description	Frame
H	Hypothesis	{H1, H2, H3}
M	Message conveyed by a card	{M1, M2, M3, M4, M5, M6, M7, M8, M9, M10, M11}
I	Information conveyed by a card	{I1, I2, I3, I4, I5, I6, I7, I8, I9, I10, I11}
I_T	Information trueness	{True; False}
Q	Source quality	{1, 2, 3, 4, 5, Unknown}
T	Source type	{AIS, LRIT, Company Security Officer, National reporting procedure, Intelligence report, Maritime Safety Agency, Smart agent, Tool providing Patterns of Life on routes, Tool providing Patterns of Life on calls, Operator + Radio, Operator + VTS + SAR}
C	Confidence	Assessed

Table 8-2: Reliability Game View*.

Variable	Description	Round 1	Round 2	Round 3	Round 4
H	Hypothesis	Assessed	Assessed	Assessed	Assessed
M	Message conveyed by a card	Provided	Provided	Provided	Provided
I	Information conveyed by a card	Provided	Provided	Provided	Provided
I_T	Information trueness	Assessed implicitly	Assessed implicitly	Assessed implicitly	Assessed implicitly
Q	Source quality	Not provided	Provided	Assessed	Provided
T	Source type	Not provided	Not provided	Provided	Provided
C	Confidence	Assessed	Assessed	Assessed	Assessed

* Assessed = player has to assess the item and communicate it to the facilitator; Provided = item value provided to the player; Not Provided = item value not provided to the player; Assessed Implicitly = player has to assess the item but not to communicate it to the facilitator.

Each card must be positioned on a game board (Figure 8-2), which is specifically designed to capture the player's belief assessment toward the different hypotheses displayed in the corners of the triangle. The selected position reflects the weight of belief that the information contained in a card provides toward some subsets of hypotheses presented. For example, positioning a card in the lower corner of the triangle indicates that the specific piece of information provided by that card is pointing towards hypothesis H1 only, while positioning the card in H1 or H2 point would indicate that the card is pointing towards both hypotheses only (i.e., excluding H3), but that the player could not discriminate between the two. The cards can be positioned not only on the points, but also on the axes connecting two points, to express some relative weight assigned to the belief towards a specific hypothesis or set of hypotheses. The shuffling of the cards during the round, after new information is discovered, is possible to allow the player's belief updating based on new evidence. Once all of the eleven cards have been processed and positioned on the board, the player is asked to rate the global confidence in the three hypotheses. For the purpose of the game, confidence was defined as "the state of feeling certain about the truth of something" [34]. The winning condition corresponds to the assignment of the highest confidence rate to the correct hypothesis. Details on the confidence rating can be found in the next section. Figure 8-3 illustrates a diagram of a game session, explaining the main actions that the participant has to perform.

To summarise, the basic GMs are:

[GM1] Assessment of hypotheses relative to a missing vessel;

[GM2] Use of cards to communicate messages to the player;

[GM3] The investigation component;

[GM4] The card positioning on the board to rate the support of a message towards hypotheses;

[GM5] The shuffling of cards as a consequence of new evidence acquisition (optional); and

[GM6] The global confidence rating of the hypotheses at the end of each round.

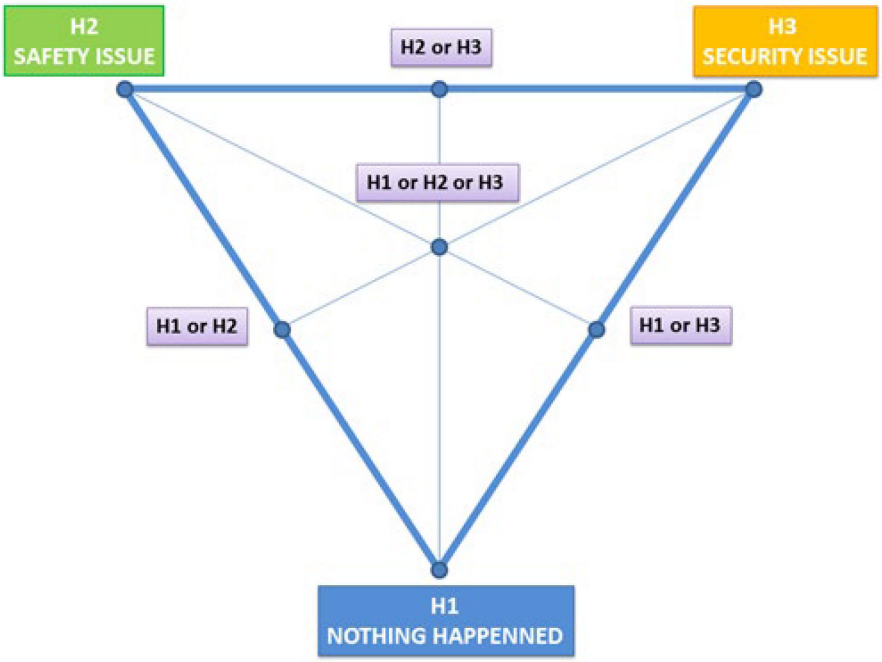


Figure 8-2: Game Board on Which the Cards Need to be Positioned.

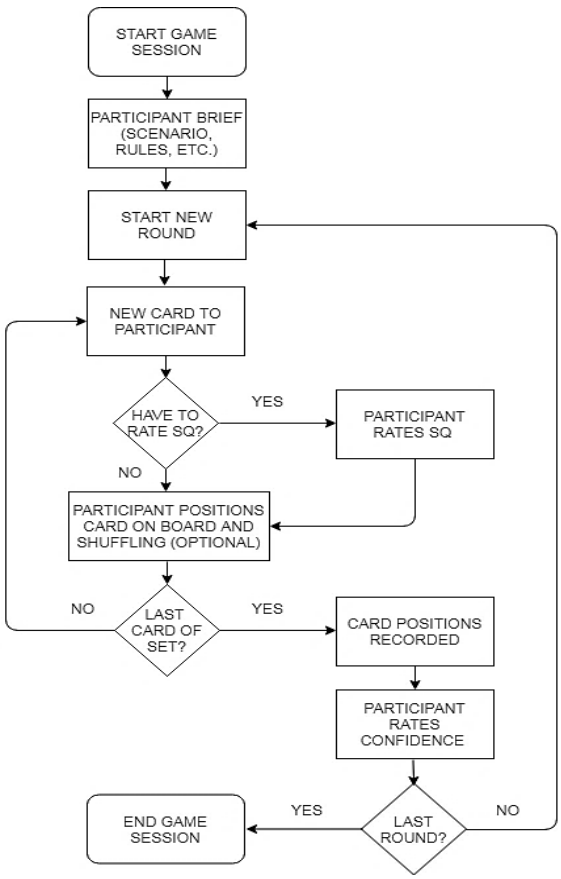


Figure 8-3: Diagram of a Session of Reliability Game.

8.4.3 Content Design

At the start of the session the player is introduced by a facilitator both to the game core and to the GMs. During this introduction session the scenario, rules and different game elements (e.g., game board, scenario map, cards and flashcards) are presented to the player. The scenario map (Figure 8-1) depicts the geographical area and other relevant geographical contextual information, such as the location of borders, the location of oil installations and the presence of a primary shipping lane that crosses the EEZ of Left Land, leading to the trans-oceanic channel. Moreover, it visualises the AIS track of the ship of interest before the contact was lost. The Messages displayed on the cards are divided in three areas; namely, the source type area, the source quality area and the information area as can be seen in Figure 8-4.

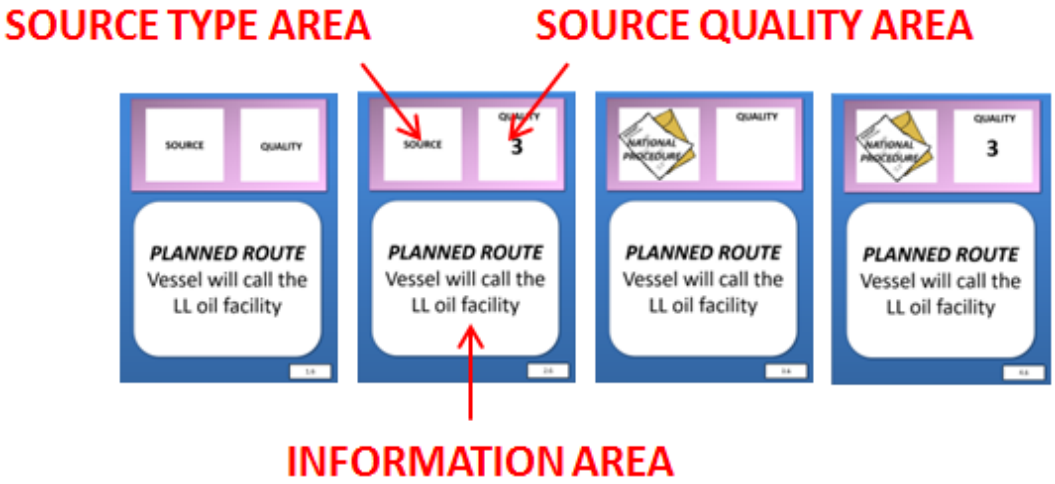


Figure 8-4: Example of the Presentation of the Same Message in the Four Different Rounds.

As previously mentioned, the only variation between the cards in the different rounds is in the meta-information on source type and source quality. As can be noted the information area is kept constant, while the source type area and source quality area are changing. Examples of message content in terms of the conveyed information, source type, source quality assigned (in Round 2 and Round 4) and information trueness is provided in Table 8-3. For instance, Message 7 reads that the Company Security Officer, which quality is unknown, reports false information regarding the fact that nothing happened to the ship.

Table 8-3: Example of Reliability Game Messages.

M	Type of Message	Source type	Source Quality Assigned	Information Trueness
1	Current ship position in X	AIS	4	False
2	Ship not answering to radio calls	Operator + Radio	5	True
7	Report that nothing is happening to the ship	Company Security Officer	U	False
8	Comparison of position X with usual ship routes	Tool providing Patterns of Life on routes	3	True

In addition to the message cards, the player is also presented with flashcards supporting the player's rating and providing additional contextual information. An example is provided in Figure 8-5, which shows the flashcards regarding the vessel of interest, the one on the source quality rating scale and finally the one on the confidence rating scale. While the first contains the relevant information regarding the ship (e.g., ship type, dimensions, flag state and ownership), the one on the source quality scale visually represents the suggested rating scale for the specific variable, which ranges from *1* to *5* or *Unknown*, with source quality *1* meaning low quality. A source quality scale of six levels has been selected to align with most of the existing standards of source reliability rating in the intelligence domain [26]. We explicitly avoided using percentage ranges and the reliability verbal expressions as studies demonstrated the subjective interpretation of the word reliability and of the matching between the verbal and numerical expression [26]. To provide an intuitive visual support to the understanding of the ranking, a graphical representation of the scale has been included, which is inspired by the home energy efficiency rating chart [35].



Figure 8-5: Flashcards – Vessel of Interest, Source Quality Levels and Confidence Levels in the Analysis.

At the end of each round, players are asked to rate the global confidence in their analysis of the current situation. The levels relative to the confidence rating are analogous to the ones for the source quality, but provide an additional definition for confidence. The rating scale was selected with reference to the analysis performed in Ref. [27]. This review shows that with respect to confidence the available rating scales differ in the proposed number of rating levels. In fact, most scales vary from a five-level scale to a three-level scale. To minimise confusion and errors, a six-level scale in agreement with the source quality scale was adopted. Those levels correspond to the five levels present in the intelligence scales plus the *Unknown* value. This value is deliberately not included in intelligence scales as it is expected that intelligence analysts are able to state their confidence in an analysis[2]. However, it was deemed worthy of inclusion in order to verify if and how this value would be used, if available. It is important to highlight that several standards defining confidence levels map the confidence terms with specific probability intervals. However, as there is no agreement on the correspondence, such a mapping was not considered in the Reliability Game.

8.4.4 Game Design Constraints

The main constraints that had to be accounted for during the design phase can be categorised as physical constraints and cognitive constraints. The first ones are those acting on the physical elements of game or related to logistical aspects, while the later are the ones dealing with cognitive tasks to be performed by the player.

The main physical constraints are the dimensions of the game elements such as the cards that had to be manageable, readable and had to be moved easily. In addition to this, another important limitation is that in

[2] Private conversation with intelligence analyst.

a non-digital game not all item moves can be easily recorded unless an external observer constantly records the moves (e.g., through notes or pictures).

The main cognitive constraints are the number of cards that have to be provided to the player, the game session length and the need for supporting elements to compensate for the fact that in real-world activities operators can rely on background knowledge and on the support of real systems (e.g., a display showing the AIS track of the lost ship).

The size of the set of cards has been selected as a tradeoff between the ability of the player to manage the set of cards and the attempt to minimise some effects that might impact the experiment results. Two notable effects are the random responding by the players [37] and the carryover effect [38]. The carryover effect takes place in within-subject experiments when one test might impact the one of the following tests. In order to minimise the carryover effect due to memorising the information from one round to the following one (also referred to as practice effect), it was decided to have a card set size major than seven. In fact, it has been suggested that the storage capacity of the short-term memory of an average person is approximately seven items, plus or minus two [39].

The game session length is a relevant cognitive constraint, as the game had to be short enough to keep the players' attention, avoiding mind-wandering effects. Mind wandering refers to the effect of the mind not focusing on a specific topic for a long period of time, which might occur especially when engaged in attention-demanding tasks [40].

8.5 GAME EVALUATION

Game evaluation is an important and critical part of design processes. In this context the term evaluation is used as proposed in Ref. [41], to avoid confusion with the concept of validation that in some contexts refers to measurements validity (e.g., measurements accuracy). Therefore, we will refer to the term evaluation as the "confirmation through the provision of objective evidence that the requirements for a specific intended use or application of a system have been fulfilled" [41], where the term "system" refers to the game to be evaluated.

As previously mentioned, the purpose of the Reliability Game is to collect data regarding source quality and source type impact on SA and SAW. Therefore, in order to evaluate the effectiveness of the game with respect to the above-mentioned scope, the main criteria are the observation of *variations of card positions* and *confidence rating* between rounds. Because the only input variation between rounds consists in the meta-information about source type and source quality, it is assumed that the two above-mentioned criteria are able to capture the corresponding impact on SA as beliefs change.

The following sections report the outcomes of a qualitative analysis of the data collected during an experiment run with the Reliability Game.

8.5.1 Experiment Set-Up

The game underwent a prototyping and play testing phase that allowed verifying the board design, the scenario, the information items proposed and the facilitation approach. After minor changes to some information items, a revised version was issued. The collection of data is still ongoing, but we present herein data collected on a small but relevant sample of SMEs that allowed verifying the effectiveness of the proposed GMs. At the time of writing this chapter, the game has been played with twenty-one (21) players, which demographics and characteristics are reported in Table 8-4. Participants' selection was performed on a voluntary base from maritime SMEs, with either civil or military status. The experimental set-up followed a within-subject design, in which the participants have been exposed to four different conditions; namely, the game rounds. The conditions variation corresponds to the game view summarised in Table 8-2.

For each player the following in-game data has been collected:

[D1] A picture of the final cards position at the end of each of the four rounds;

[D2] The source quality rating during the third round; and

[D3] The confidence rating in the hypothesis at the end of each round.

Table 8-4: Participant Demographics and Characteristics.

Feature		
Gender	Male	100%
	Female	0%
Age	Average	46.5 years
	Standard Deviation	10.3 years
Status	Military	76%
	Civilian	24%
Nationality	Italian	33.33%
	French	4.76%
	Danish	14.28%
	Norwegian	4.76%
	Romanian	4.76%
	British	14.28%
	German	19.04%
	American	4.76%

In this experiment there was not an external observer constantly recording the item movements. Thus, only the final aggregation of beliefs at the end of each round has been recorded (D1) (see Figure 8-6), while the shuffling of cards has not been captured. However, this represents a minor issue as the cards shuffling resulted in a kind of GMs seldom used by the players. Moreover, it will be completely superseded in a digital version of the game, currently under development.

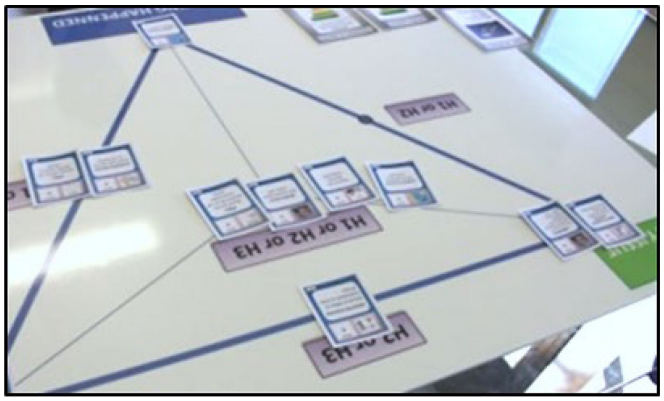


Figure 8-6: Example of a Picture Collected at the End of a Round (D1).

Beside the in-game data collection mentioned above, a post-game data collection has been performed in the form of feedback questionnaire. The scope of this questionnaire was to assess participants' understanding of the game and perception with respect of this innovative gaming approach (e.g., relevance with respect to their mission, engagement, facilitation). It is important to note that this questionnaire has been provided as part of a broader feedback questionnaire and only 11 out of the 21 players of the Reliability Game returned their answers.

8.5.2 Feedback and Observations on the Game Design

The participant survey shows that the players perceived the game as engaging, realistic and relevant with respect to operational needs (Figure 8-7). From a facilitation point of view, it has been observed that it is important not only to introduce the players to the game rules and to have them familiar with the game dynamics, but also to clearly state and explain the game core to have the players feeling more comfortable and confident about the remaining part of the experiment. Most players actually were explaining their reasoning to the facilitator, which is considered of value for the refinement of next iterations of the game.

Players showed good understanding of the purpose of the game and the GMs, which appears to be intuitive and requires a low level of pre-experiment training. Note that there is not a proper pre-experiment training session. Instead, the rules are explained and then the facilitator guides the player when providing the first cards by asking after the card is positioned if the player confirms that the card supports the belief associated to the specific card position. In case of a negative answer, the facilitator would help the player positioning the card in the corresponding location.

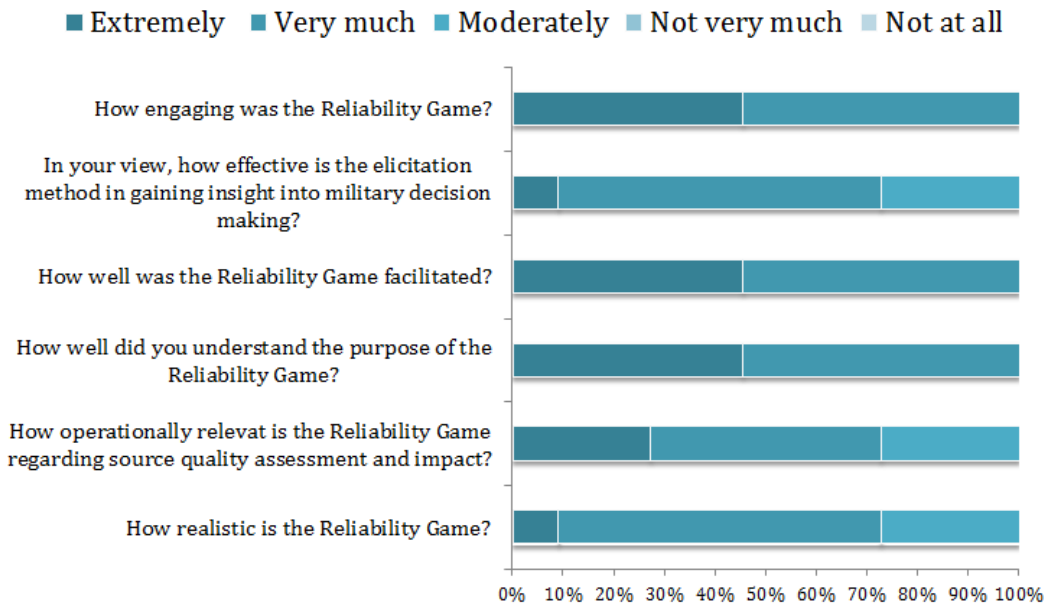


Figure 8-7: Players' Feedback Questionnaire Outcomes.

8.5.3 Outcomes on Source Quality Rating

During the third round, of the game participants were requested to rate the source quality given the meta-information on source type which had been provided (Section 8.4.3). Figure 8-8 presents an example of the source quality rating by three different players (red diamonds), which is compared to the source quality values that are provided to the players in Round 2 and Round 4 (blue line). Empty values correspond to an *Unknown* rating. The three players presented different rating profiles. We can observe how Player A has a

tendency to rate the source quality higher than the assigned source quality value. Player B demonstrates a tendency to variably rate the source quality higher, lower or equal to the assigned ratings. Finally, Player C shows a tendency to rate the source quality lower than the quality assigned.

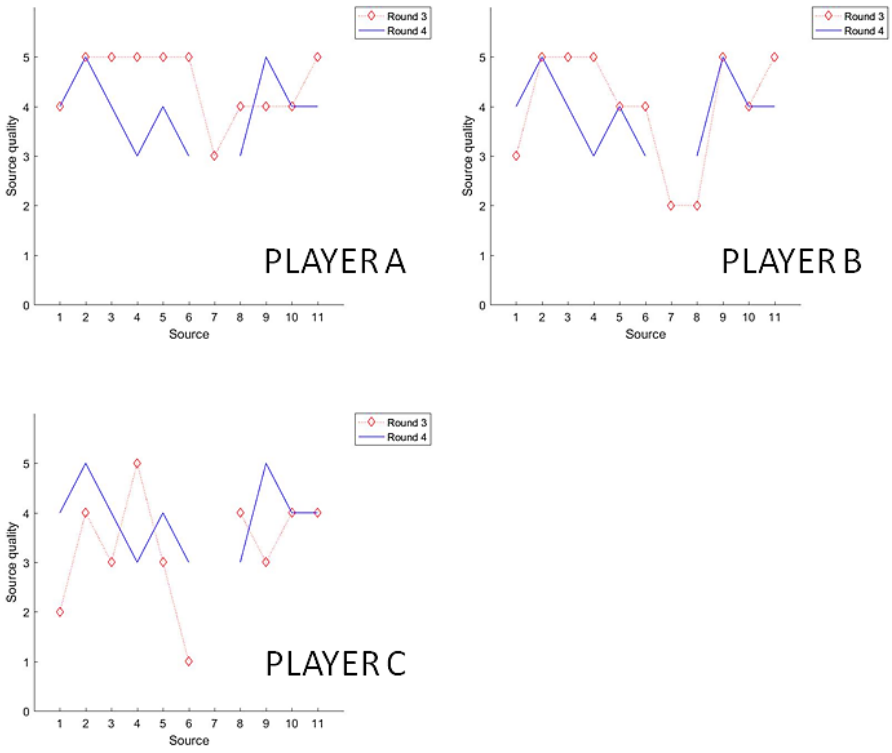


Figure 8-8: Example of Source Quality Rating (Round 3) by Three Different Players.

Figure 8-9 depicts the overall source quality assessment for each of the 11 cards, as a percentage of players assigning a given source quality rating by card. Note that although the source quality values with decimals (*3.5* and *4.5*) were not included in the original scale, one player requested to use them. From the figure it can be observed that with the exception of the rating of this player, the assessments on Source 1 and Source 8 are identical. This is an important observation because the degree of familiarity of the SMEs with the two sources is considerably different. In fact, Source 1 (Automatic Identification System) is widely available and commonly used in maritime surveillance. On the contrary, Source 8 (Vessel to Route Association algorithm) is more experimental and is still in its early stages of development. Other novel information sources are Source 5 and Source 10; namely, a vessel position prediction algorithm and a maritime Patterns of Life on ship statistics service. Both sources present a certain degree of variation in the quality rating. However, the result suggests that non-conventional information sources are not necessarily considered of low quality. Moreover, from the players' verbal feedback it appeared that the players were drawing comparisons between the source's capacity and their own cognitive abilities (e.g., ability of associating a ship to a route). This observation concurs with some persuasion literature on source factors which has shown the impact of the perceived source similarity on human information assessment [29]. Source 7 (Company Security Officer) is the one exhibiting the highest degree of variability in the quality ratings. This source in Round 2 and Round 4 has an *Unknown* assigned quality. Only three players rated the source as such, while most of the players assigned a low-quality rating. As explicitly stated by some players, this appears not to be related to the nature of the source (human vs. sensor), but rather to a possible conflict of interest of the Company Security Officer who could retain or falsify information due to conflict of interests. This observation is also supported by the fact that Source 2, the human operator, has been rated of high quality.

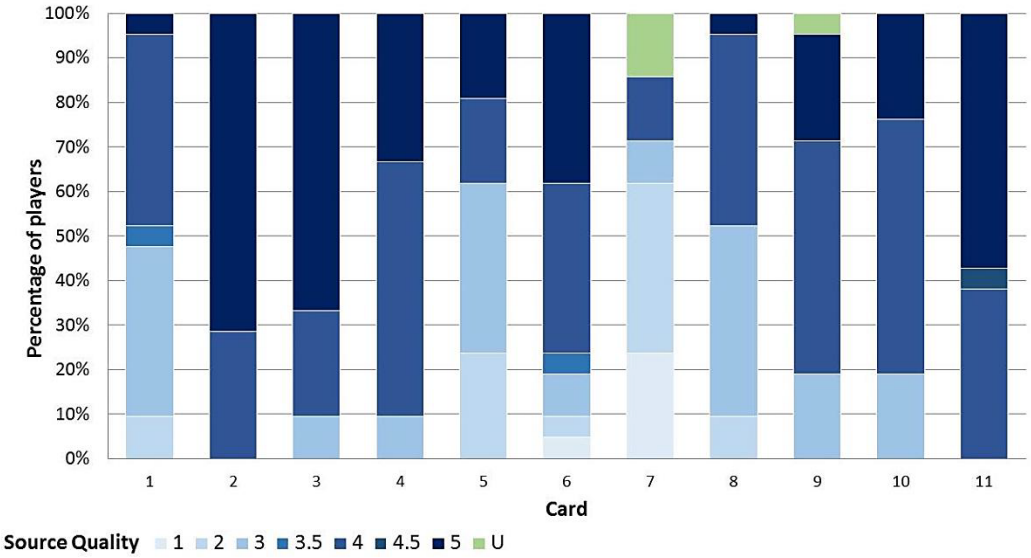


Figure 8-9: Source Quality Ratings by Card.

An interesting result is the one related to the use of the value *Unknown*. In fact, it has been seldom used, even in the case in which the player had no knowledge of the type of source. More specifically, some players did not know the Company Security Officer or the Long-Range Identification and Tracking system. They asked for information to the facilitator, who provided basic information, without disclosing details on the quality. Although the players were often reminded of the possibility of using the *Unknown* value, most of them did not do. This suggests that the players tended to estimate a source quality value even if the source is not known.

Moreover, as shown in Figure 8-8 there was a difference between the participants' source quality ratings and the values provided to them in Round 2 and Round 4. This translates in a variation of the conditions between Round 3 (source type provided, source quality assessed) and Round 4 (source type provided, source quality provided).

8.5.4 Outcomes on Confidence Rating

At the end of each round, participants were requested to rate their confidence in the fact that the correct hypothesis might be H1, H2 or H3. Figure 8-10 displays the confidence rating of different players. With respect to the relative confidence ratings, it can be observed that the sum of the confidence in the hypotheses is not constant between the rounds and that the variation of the confidence in one of the hypotheses does not imply the variation of the confidence in the others.

From Figure 8-11, which is reporting a summary of the different confidence ratings of the participants, we can observe interesting results regarding the use of the scale presented. Equivalently to the case of source quality, one player asked to use a value with decimals. More specifically, the participant asked to introduce the value *2.5* as to express the concept of *50%*, which is not possible with the original form of the scale. The proposed scale did not include the rating value *0*, which conceptually corresponds to the exclusion of the hypothesis with high confidence. However, many players asked to use the value *0*. On the contrary the value *5*, corresponding to the conceptual opposite (full confidence that the specific hypothesis is the right one), has been rarely used. This result suggests that participants might more easily exclude hypotheses than confirming hypotheses. Another possible interpretation is that they might feel more self-confident in excluding than being certain about the hypotheses. With the term self-confidence the authors refer to the concept of self-assurance in personal judgement.

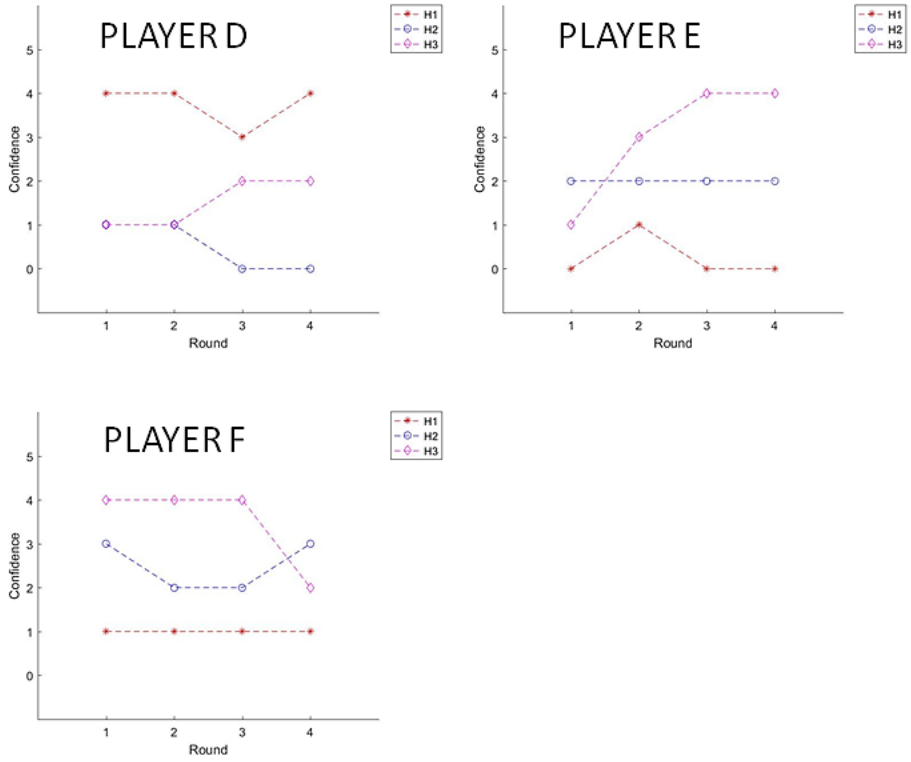


Figure 8-10: Example of Confidence Rating by Different Players.

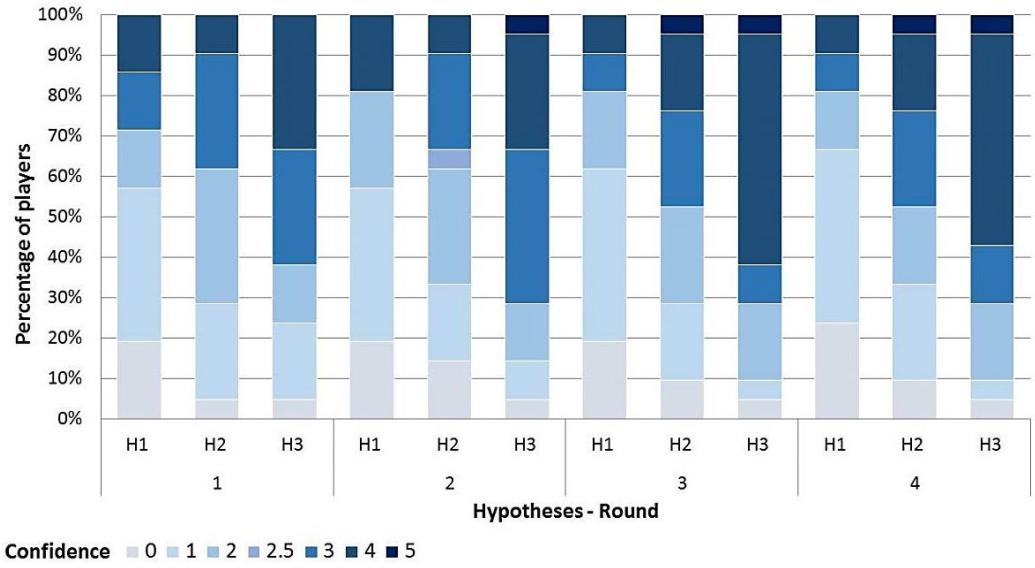


Figure 8-11: Confidence Ratings by Hypothesis in the Different Rounds.

Contrary to the intelligence scales, a sixth level was included in the original scale (i.e., *Unknown*), to allow the players the possibility to state their inability to draw a conclusion and express their confidence. This value is intentionally excluded from the intelligence standard scales as it forces the analyst to exactly rate his confidence, no matter if high or low, without using the aforementioned value as a solution to avoid liability issues. It is, however, noteworthy that this value has been used twice by the participants, but this is not reported in the graph as the players soon after asked if they could re-rate the confidence.

Another interesting observation regarding the confidence rating is that some players when requested to express their confidence were stating that it was unchanged with respect to the previous round. The facilitator, however, requested the players to explicitly rate the current confidence levels. The resulting rating, in general, was not equal to the previous one, suggesting that there had been a change of which the players were not conscious. This also suggests that the mechanisms to try to minimise the carryover effect were effective on the confidence rating.

8.5.5 Outcomes on Card Positions

At the end of each round, a picture of the board has been taken. From the pictures the authors have been able to record all the data regarding the single card assessment. Table 8-1 reports an example of the card positions in the different rounds played by one participant. Note that although this example shows only cards positioned in the points specified on the board (corners, mid of axes and centre of the triangle) the players, in general, used the full spectrum of the possible positions; more specifically, the axes displayed on the board (Figure 8-2).

Table 8-5 shows how position variations between the rounds have been consistently observed. This table reports only the final position for each round, while the card shuffling is not reported. This is because although the players have been allowed to shuffle cards during the game (GM6), this GM has been used only twice during the experiment run.

Table 8-5: Example of Card Positions Collected for Each Player.

Card	Position Round 1	Position Round 2	Position Round 3	Position Round 4
1	H1	H1	H3	H3
2	H1 or H3	H1	H1	H3
3	H2	H1 or H2	H1 or H2	H1 or H2
4	H3	H1 or H3	H3	H1 or H3
5	H2	H3	H1	H1
6	H2	H1 or H3	H1 or H3	H1 or H3
7	H1 or H2 or H3	H1 or H2 or H3	H1	H1 or H2 or H3
8	H3	H1 or H3	H1 or H3	H1
9	H3	H1 or H2 or H3	H1 or H2 or H3	H3
10	H3	H3	H3	H3
11	H3	H3	H3	H3

From a qualitative analysis, it has been possible to observe the impact of source quality and source type on players' assessments by means of the change of the card positions on the board between the different rounds played by the same participant. However, a more in-depth analysis is required to be able to quantify this impact and draw connections between those factors, the player SA and final confidence. Such an analysis requires a formalisation of belief assessment together with a proper encoding of the players' cards positions. The mathematical framework should be rich enough to capture the uncertainty expressed by the players and conveyed through the board game. Evidence theory [41] will be the favoured framework for its ability to express total ignorance and non-additive assessments.

8.6 CONCLUSION

To take full advantage of the variety of information within systems that support SA, the underlying fusion processes should properly account for source factors. In order to enable this capability, research is still required with respect to the characterisation and quantification of those factors on SA. To this end, we developed a data exchange game, called the Reliability Game. The purpose of this game is to collect data regarding players' belief changes as a function of source factors – more specifically, source type and quality. To gather such data each player is presented with a scenario and plays several rounds of the game. The only variation between rounds consists in his knowledge regarding source type and quality. The corresponding belief changes are captured though the variation of game items position (cards) and final confidence ratings.

We performed a qualitative analysis on the data gathered through an experiment run with the Reliability Game in order to evaluate the effectiveness of the game design and GMs. Moreover, we performed a post-game data collection in the form of a feedback questionnaire. The results show that the game is perceived both as engaging and relevant. Moreover, the game scope and game mechanics were easily understood.

The variations of the players' belief assessments between the different rounds demonstrate that the proposed methodology effectively captures elements of source factors impact on SA. Therefore, the Reliability Game is an innovative method that allows gaining insight into the definition of the reliability underpinning factors. Moreover, the analysis demonstrates the capability of the game to highlight important aspects of the use of the rating scales. This makes the Reliability Game a powerful tool to collect data relevant to the standardisation efforts in communication of uncertainty (e.g., confidence, reliability, credibility).

8.7 REFERENCES

[1] IGI Global, "What is serious games?", Retrieved from https://www.igi-global.com/dictionary/serious-games/26549.

[2] Carvalho, M.B., Bellotti, F., Berta, R., De Gloria, A., Sedano, C.I., Baalsrud Hauge, J., Hu, J., and Rauterberg, M. (2015). An activity theory-based model for serious games analysis and conceptual design. *Computers and Education* 87:166-181.

[3] Roungas, B. (2016). A model-driven framework for educational game design. *International Journal of Serious Games,*3(3).

[4] Strzalkowski, T., Symborski, C. (2016). Lessons learned about serious game design and development. *Game and Culture,* 12(3):292-298.

[5] Vermillion, S., Malak, R., Smallman, R., Becker, B., Sferra, M., and Fields, S. (2017). An investigation on using serious gaming to study human decision-making in engineering contexts. *Design Science* 3 (E15).

[6] Djaouti, D., Alvarez, J., and Jessel, J.-P. (2011). Classifying serious games: The G/P/S model. In: *Handbook of Research on Improving Learning and Motivation through Educational Games: Multidisciplinary Approaches*, 1:118-136. Portland, OR: Ringgold Inc.

[7] Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popović, Z. (2010). Predicting protein structures with a multiplayer online game. *Nature,* 466(7307):756-760.

[8] Ren, Y., Bayrak, A.E., and Papalambros, P.Y. (2016). EcoRacer: Game-based optimal electric vehicle design and driver control using human players. *Journal of Mechanical Design,* 138(6):061407-1–061407-10.

[9] Endsley, R.M. (1987). The application of human factors to the development of expert systems for advanced cockpits. In: *Proceedings of the Human Factors Society 31st Annual Meeting*, pp. 1388-1392. Human Factor Society, Santa Monica, CA.

[10] Graafland, M., and Schijven, M.A. (2013). A serious game to improve situation awareness in laparoscopic surgery. In: *Games for Health*, Schouten, B., Fedtke, S., Bekker, T., Schijven, M., and Gekker, A. (Eds.), 173-182. Wiesbaden, Germany: Springer Vieweg Verlag.

[11] Sawaragi, T., Fujii, K., Horiguchi, Y., and Nakanishi, H. (2016). Analysis of team situation awareness using serious game and constructive model-based simulation. In: *Proceedings of the 13th IFAC Symposium on Analysis, Design, and Evaluation of Human-Machine Systems HMS 2016*, Elsvier, 49(19): 537-542.

[12] de Rosa, F., Jousselme, A.-L., and De Gloria, A. (2018). A reliability game for source factors and situational awareness experimentation. *International Journal of Serious Games* 5 (2):45-64. https://doi.org/10.17083/ijsg.v5i2.243.

[13] Endsley, R.M. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors* , 37(1):32-64.

[14] Stikeleather, J. (2012). Big data's human component. *Harvard Business Review.* Retrieved from https://hbr.org/2012/09/big-datas-human-component%20retrieved%20on%2010/10/2017

[15] Endsley, R.M. (1996). Automation and situation awareness. In: *Automation and Human Performance: Theory and Applications*, Human Factors in Transportation Series, Parasuraman, R., and Mouloua, M. (Eds.), 163-181, Hillsdale, NJ: Lawrence Erlbaum Associates.

[16] Pew, R. (1985). Human skills and their utilization. In: *Human Factors Engineering: Engineering Summer Conferences*. University of Michigan, Ann Arbor, MI.

[17] Darlington, K., Explainable AI systems: Understanding the decisions of the machines, BBVA OpenMind, 2017. Retrieved from https://www.bbvaopenmind.com/en/technology/artificial-intelligence/explainable-ai-systems-understanding-the-decisions-of-the-machines/#.WkEZgAJbI-A.twitter

[18] Hall, D.L., and Jordan, J.M. (2010). *Human-Centered Information Fusion*. Boston, MA: Artech House.

[19] Wiener, E.L. (1988). Cockpit automation. In: *Human Factors in Aviation (Cognition and Perception),* In Wiener, E. L., Nagel, D. C. (Eds.), 433-461. San Diego, CA: Academic Press Inc.

[20] Marusich, L.R., Bakdash, J.Z., Onal, E., Yu, M.S., Schaffer, J., O'Donovan, J., Höllerer, T., Buchler, N., and Gonzalez, C. (2016). Effects of information availability on command-and-control decision making: Performance, trust, and situation awareness. *Human Factors,* 58(2):301-321.

[21] Rogova, G., and Nimier, V. (2004). Reliability in information fusion: Literature survey. In: *Proceedings of the 7th International Conference on Information Fusion,* 1158-1165. Stockholm, Sweden.

[22] Pichon, F., Mercier, D., Lefèvre, É., and Delmotte, F. (2016). Proposition and learning of some belief function contextual correction mechanisms. *International Journal of Approximate Reasoning,* 72:4-42.

[23] Pichon, F., Dubois, D., and Denoeux, T. (2012). Relevance and truthfulness in information correction and fusion. *International Journal of Approximate Reasoning,* 53:159-175.

[24] Haenni, R., and Hartmann, S. (2006). Modeling partially reliable information sources: A general approach based on Dempster-Shafer theory. *Journal of Information Fusion,* 7 (4):361-379.

[25] Stanton, N.A., Salmon, P.M., Walker, G.H., Baber, C., Jenkins, D.P. (2006). *Human Factors Methods: A Practical Guide for Engineering and Design*. Burlington, VT: Ashgate Publishing Company.

[26] De Rosa, F., Jousselme, A.-L., and De Gloria, A. (2017). Gamified approach in the context of situational assessment: A comparison of human factors methods. In: *Advances in Human Factors, Software, and Systems Engineering, Proceedings of the 9th International Conference on Applied Human Factors and Ergonomics*, 100-110. New York, NY: Springer.

[27] De Rosa, F., and Jousselme, A.-L. (2019). Critical review of uncertainty communication standards in support to maritime situational awareness. Technical Report. CMRE-FR-2018-010, NATO STO Centre for Maritime Research and Experimentation. La Spezia, Italy.

[28] Lexico, "Reliability", (n.d.), Retrieved from https://www.lexico.com/definition/reliability.

[29] Briñol, P., and Petty, R.E. (2009). Source factors in persuasion: A self-validation approach. *European Review of Social Psychology,* 20(1):49-96.

[30] Brathwaite, B., and Schreiber, I. (2008). *Challenges for Game Designers*. Charles River Media. Rockland, MA.

[31] Jousselme, A.-L., Pallotta, G., and Locke, J. (2015). *A risk game to study the impact of information quality on human threat assessment and decision making*. Technical Report CMRE-FR-2015-009, NATO STO Centre for Maritime Research and Experimentation, La Spezia, Italy.

[32] International Organization for Standardization. (2011). *ISO 5725-1:1994(en), Accuracy (Trueness and Precision) of Measurement Methods and Results* – Part 1: General principles and definitions. Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:ed-1:v1:en

[33] Yankova, K. (2015). *The Influence of Information Order Effects and Trait Professional Skepticism on Auditors' Belief Revisions*. Wiesbaden, Germany: Springer Gabler.

[34] Lexico, "Confidence", (n.d.), Retrieved from https://www.lexico.com/definition/confidence.

[35] Wikipedia. Energy Performance Certificate (United Kingdom)", Retrieved from https://en.wikipedia.org/wiki/Energy_Performance_Certificate_(United_Kingdom).

[36] Osborne, J.W., and Blanchard, M.R. (2010). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology*,1.

[37] Brooks, J.L. (2012). Counterbalancing for serial order carryover effects in experimental condition orders. *Psychological Methods,* 17(4):600-614.

[38] Millar, A.G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 101 (2):343-352.

[39] McVay, J.C., and Kane, M.J. (2009). Conducting the train of thought: Working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology, Learning, Memory, and Cognition,* 35(1):196-204.

[40] Wilson, D.W., Jenkins, J., Twyman, N., Jensen, M., Valacich, J., Dunbar, N., Wilson, S., Miller, C., Adame, B., Lee, Y.-H., Burgoon, J., and Nunamaker, J.F. (2016). Serious games: An evaluation framework and case study. In: *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS),* Koloa, HI.

[41] Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton University Press.

Chapter 9 – THE RISK GAME: CAPTURING IMPACT OF INFORMATION QUALITY ON HUMAN BELIEF ASSESSMENT AND DECISION MAKING[1]

Anne-Laure Jousselme, Giuliana Pallotta, and Jonathan Locke
NATO STO Centre for Maritime Research and Experimentation (CMRE)
ITALY

9.1 INTRODUCTION

Reaching an adequate level of Situation Awareness (SAW) for an informed and confident decision not only requires processing information of various types (numerical values, natural language statements, objective or subjective assessments, etc.) and dealing with imperfect information (uncertain, imprecise, conflicting, doubtful, ambiguous, etc.) but requires also understanding and taking advantage of the operational context. Moreover, it requires processing information across different levels of semantic content, ranging from lower levels of processing (e.g., contact and target detection, target tracking, classification) to higher levels of processing (e.g., activity recognition, behaviour analysis, target identification, threat and impact assessment), while considering a user's needs and context [2]. Processing information which can be irrelevant and thus distracting is highly demanding for the human operator who is often under stress due to time constraints. All these factors have a measurable impact on the decision in terms of timeliness, confidence or correctness.

The impact of information quality on decision quality has been studied in different domains such as finance [3] or enterprise performance [4], [5]. Within these works, different information quality dimensions are considered and their impact on the decision maker is highlighted. For instance, the quality dimensions of *accessibility*, *accuracy*, *completeness*, *interpretability* are used in Ref. [5]. Only a few works study the impact of information quality on situation and threat assessment in the maritime domain, but it is worth mentioning the *Data Fusion Levels Two and Three Workshop* sponsored by the Office of Naval Research (ONR) and held in Arlington, VA in 2005. The summary paper [6] describes "how information pedigree is used to support and enhance situation and threat assessment", considering dimensions such as *reliability*, *confidence*, *trust*, and their integration in automation of sensor-data processing. In Ref. [5], the authors follow Wang's Total Data Quality Management (TDQM) cycle [7], which implements four components:

1) Identification of information quality dimensions;
2) Measurement, providing information quality metrics;
3) Analysis, identifying causes for problem and computing the impacts of poor information quality; and
4) Improvement, providing techniques for improving information quality.

The approach proposed in this current chapter follows the steps of Ref. [7] and is comparable to the one proposed in Ref. [4], while the impact of information quality on decision quality is studied along the quality dimensions of *falseness*, *imprecision*, *uncertainty* and *relevance*.

Such studies rely on different methodologies: for instance, in Ref. [8], semi-structured interviews of Marine Protected Area (MPA) experts are conducted and their interpretation of different attributes is analysed. Hoffman *et al.* [9] distinguish between three categories of elicitation techniques: "1) Analysis of the tasks that experts usually perform; 2) Various types of interviews; and 3) Contrived tasks which reveal an expert's reasoning processes without necessarily asking about these processes". The Risk Game

[1] An extended version of this chapter is available as a NATO CMRE report by Jousselme et al. [1].

presented here falls under the third category. Using games for Knowledge Elicitation (KE) has the advantage over structured or semi-structured interviews (e.g., Ref. [8]), of being more engaging for the experts and less time consuming [9], [10]. The effectiveness of serious games to enhance the risk management process has been recently reported in Ref. [11]. A study published in 2018 [12] describes a serious game for natural risk assessment where an interactive interface is designed and some indicators (such as motivation and immersion) are used to assess the players' feedback. A very recent survey [13] provides an interesting discussion on serious games role in learning disaster risk management. Serious games are also an efficient means to highlight reasoning biases as reported, for instance, in Refs. [14] and [15].

The main research question addressed in this study is whether serious games are an efficient means to capture the impact of information quality on human belief assessment and decision making. A secondary question is which information quality dimension impacts belief assessment. To this end, we hypothesise that information quality can be reduced to the three basic information quality dimensions of uncertainty, imprecision and falseness. We also assume that human belief assessment is impacted by the type of the source of information and the nature of information (i.e., attribute), while the decision depends on the operational context and associated risk.

The Risk Game presented in this chapter is a gaming approach to elicit experts' knowledge and know-how in processing heterogeneous information (from sensor measurements to human statements), considering information and source quality and reasoning about concurrent events. It is a contrived technique in the sense of Ref. [9], analysing the quality dimensions of uncertainty, imprecision and falseness. Taking a "game with a purpose" approach [16], the Risk Game is aimed at capturing data expressing human reasoning capabilities while performing a specific task of maritime situation assessment, but can be adapted to any other application domain. Further analysis of the structured data gathered during experiments would contribute to the development of automated algorithms for an improved synergy with the human operator. The scenario-based design approach makes it easy to develop several versions of the games suited to particular applications. In this chapter, the maritime surveillance problem is tackled, but intelligence applications can be considered as well, in support to the intelligence analyst hypotheses evaluation task.

The chapter is organised as follows: In Section 9.2 we present the Risk Game design and detail the methodology developed. Section 9.3 presents the players' feedback gathered through a survey at the end of the exercise. The experimental setting involving the players together with some exploratory data analysis is presented in Section 9.4. Section 9.5 summarises some findings and sketches some research avenues to be further explored.

9.2 GAME DESIGN

The Risk Game has been designed to validate the hypothesis that Information Quality (IQ) has an impact on Belief Assessment (BA), Risk Assessment and Decision Quality. We present in this section a methodology which allows:

1) To elicit experts' knowledge and know-how about belief and risk assessment and decision making, in an entertaining way;
2) To gather data which once analysed, would support or refute our hypothesis; and
3) To characterise and measure such an impact.

9.2.1 Game-Play

A set of players has first been selected, mostly experts in maritime surveillance and threat assessment, who played individually and successively (see Figure 9-1) following the chronology of the game round:

- The player plays the role of the Officer of the Watch (OoW) who has to assess a vessel behaviour which track has been lost one hour ago.
- The player is presented two candidate tracks (corresponding to Vessel *A* and Vessel *B*) that could correspond to the lost vessel, with approximate positional information: One track is located in the Area Of Responsibility (AOR), the other one being in the country of origin of the lost vessel. The player does not know that that in the scenario of the game, Vessel *A* is the missing vessel and thus should be assessed as a threat. Vessel *B* is a fishing vessel from outside the AOR which is going back to its port. The added difficulty in the threat assessment is that the two vessels are of the same type.
- The player should follow the following general reasoning: If the track in the AOR corresponds to the lost vessel, then it means that the lost vessel crossed the border without authorisation and should be considered as a potential threat; however, if the track not in the AOR corresponds to the lost vessel, then that means that the lost vessel is still in its area and there is no threat.
- The player successively selects Pieces of Information (PoIs) of randomized quality, about the two observed and unknown tracks to discover which one corresponds to the lost vessel.
- After discovering each PoI, the player is asked to assess which of the two tracks corresponds to the lost vessel, under the form of a belief degree, and fill the corresponding assessment form accordingly.
- At any time during the Risk Game (i.e., after having examined an arbitrary PoIs), the player can take the final resolution to either send or not to send a patrol aircraft to inspect the vessel. This decision marks the end of the game. The ultimate goal of the Risk Game is to take the best decision about a threat by using the minimum number of PoIs.



Figure 9-1: A Player Assessing the Threat from an Unknown Track in the Vicinity of the Port, Based on Information Previously Queried.

The game mechanics are articulated around on three main elements:

1) A scenario describing a story and settling the player in a specific role with specific task in a specific context;
2) Information cards abstracting pieces of information which are familiar to the domain experts and usually displayed on screens or providers by operators; and
3) A belief gathering method to capture experts' sequentially updated belief degrees.

The independent variables cover a wide range of information quality dimensions among which three (falseness, uncertainty and imprecision) are controlled and randomized by dice roll. Their impact on experts' belief degrees and decision made (dependent variables) are captured.

The board version of the Risk Game presented in this chapter is attractive and playful thanks to:

1) A coloured board and a set of coloured cards which allows a nice visual discovery phase of the game;
2) The actions from the player of rolling the dice, selecting cards and rating the belief degrees which keep the player engaged;
3) The sometimes-challenging reasoning task which forces the player to maintain the focus on the task; and
4) An interactive facilitation which makes the dialogue with the scientist easy and makes the sharing of experience and knowledge pleasant and efficient.

9.2.2 Scenario-Based Game Design

A short scenario description with some context is given to the players, who will be actors in the story, with the role of decision maker. The vignette developed is part of the larger general scenario which takes place in the port of Herosé (Figure 9-2(a)) at the border between Centre Land (the AOR) and Right Land (to which the lost fishing vessel belongs) after a period of crisis, and the Harbour Protection Level is set to TWO (over a scale of three levels). Information about the lost vessel is provided to the player as would be the case from typical vessel databases: Name, Type, Subtype, Length, Width and Flag, together with its last Location, Heading and Speed (Figure 9-2(b)).

9.2.2.1 Underlying Reasoning

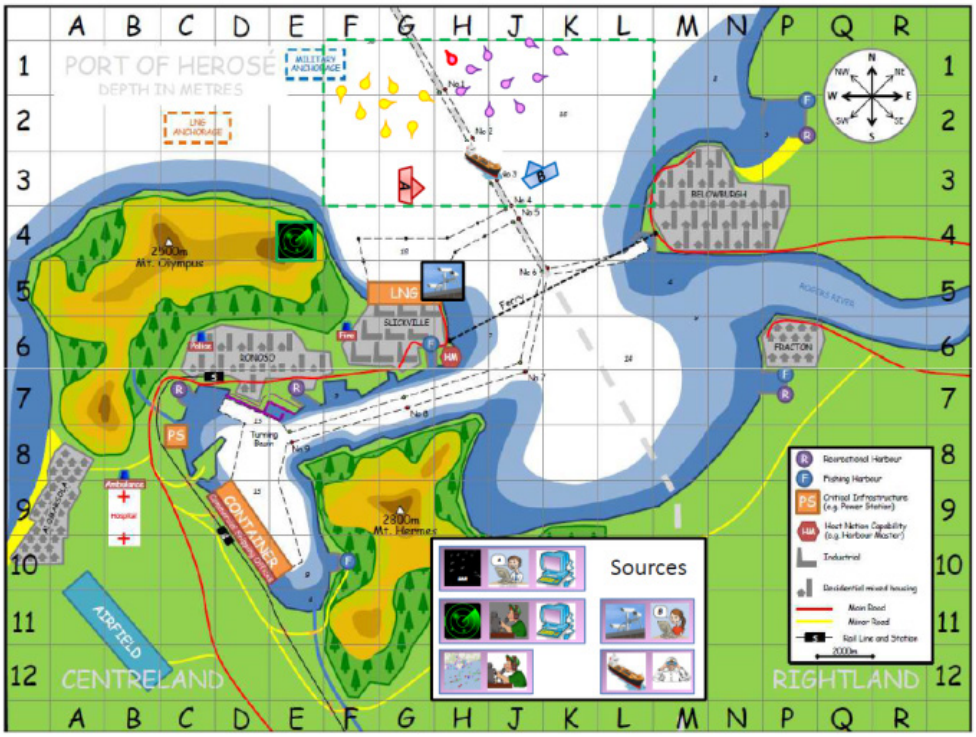
Let us denote by A the event (or proposition) "The lost vessel corresponds to Track A" (in Centre Land) and by B the event "The lost vessel corresponds to Track B" (in Right Land). The two events are exclusive since the lost vessel cannot be in both Centre Land (event A) and Right Land (event B). Thus:

$$A \wedge B \vdash \bot \tag{9-1}$$

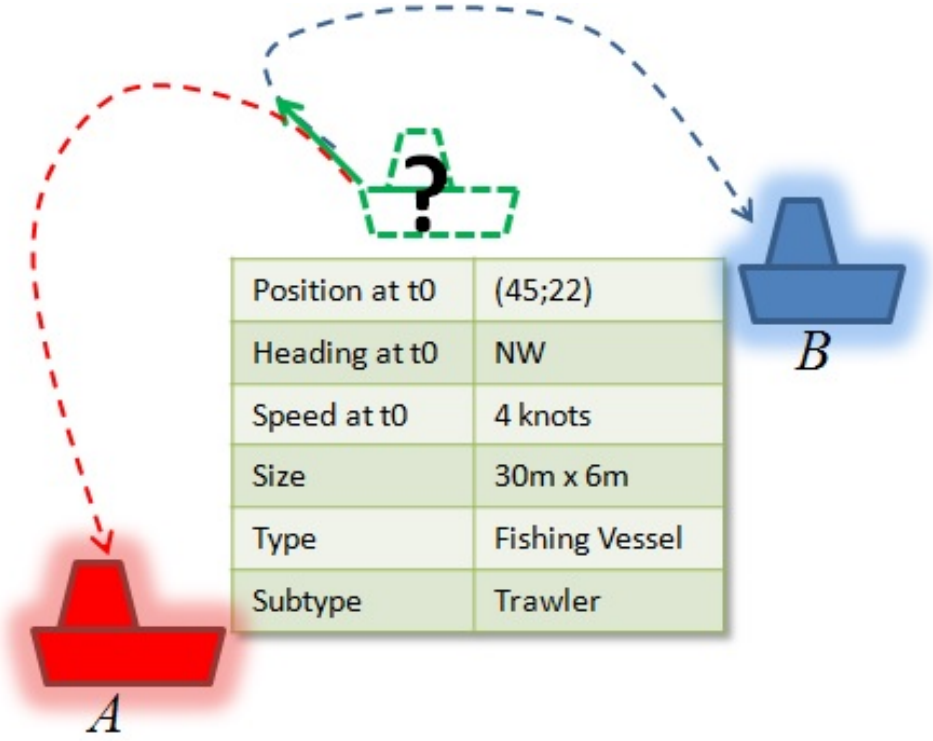
where $\wedge$ denotes the logical conjunction, and $\bot$ denotes the logical contradiction. Equation 9-1 reads: "Event A AND Event B cannot be both true". We however do not exclude the possibility that neither A nor B is true meaning that the vessel may have left the area under surveillance. The exhaustivity of the events is referred to in Ref. [17] as the *open-world* assumption and writes:

$$A \vee B \nvdash \top \tag{9-2}$$

where $\top$ is the logical tautology.



(a) Port of Herosé, Where the Scenario Takes Place.



(b) The Lost Vessel Corresponds Either to Track *A* (Red) or Track *B* (Blue).

Figure 9-2: Board and Scenario for the Risk Game.

Any evidence for *A* should either at worst decrease the belief for *B*, i.e., provide some evidence that the lost vessel is Vessel *B,* or at least should not provide any information. Equivalently, any evidence toward *B* should either at worst decrease the belief for *A*, or at least should not provide any information. For instance, if the player learns (based on some information the player is provided with) that Track *B* does not correspond to a fishing vessel, the player's belief that Track *A* is the lost vessel should increase. Let us denote by Bel(*A*) the belief degree assigned by the player toward the event *A*. It is known that human belief assessment does not necessarily satisfy probabilistic coherence (see, for instance, Ref. [18]), meaning that the belief degrees towards events *A* and *B* may not sum up to 1: They can be lower than, higher than or equal to 1. In the case the player follows a probabilistic assessment, the coherence principle under the *closed-world* assumption imposes that Bel(*A*)+Bel(*B*)=1. Under the open-world assumption, the same principle would be stated as Bel(*A*)+Bel(*B*)≠1 meaning that another event not considered here (i.e., the vessel is outside the area) would be assigned the complementary degree to 1.

To respect the natural non-additive belief assessment of humans while considering both open- and closed-world assumptions, we allow the player a complete freedom in belief degree assessments about the two events. The belief degrees elicited about events *A* and *B* are thus free from any probabilistic interpretation and thus do not need to sum up to 1.

9.2.2.2 Other Applications

The scenario-based game design allows the flexibility to customise the game for specific applications. In particular, the problem of aggregating (i.e., fusing) pieces of information fusion under uncertainty appears not only in maritime situation awareness applications, but also in intelligence applications where the aggregation of experts' opinions expressed with uncertainty is rather known as "multi-INT fusion". Some methods such as the Analysis of Competing Hypotheses (ACH) are aimed at supporting the intelligence analyst in evaluating multiple hypotheses provided by different sources, considering some quality dimensions such as credibility or relevance. The methodology proposed in this chapter is flexible enough to design a version of the Risk Game suited to the multi-INT problem. The key step is the choice of a relevant scenario supported by a set of meaningful sources of information.

9.2.3 Attributes

Each vessel is described by a series of five attributes: LOCATION, SPEED, HEADING, TYPE, SIZE. These attributes have been selected to cover different types of scales (continuous, discrete, nominal, etc.). Moreover, some are independent (e.g., LOCATION and SPEED) while others are dependent (e.g., SIZE and TYPE).

We define the *a priori* effective relevance[2] of information regarding the attribute only, ending up with the qualitative ranking and further categorise the attributes as behavioural: LOCATION, SPEED, HEADING and classification: TYPE, SIZE, SPEED. The task of the player is to discover "where" the suspect vessel is, or other said which track it corresponds to. It is thus essentially a classification problem where the player needs to *match* queried information with the corresponding attributes know about the lost vessel. The behavioural attributes are thus less relevant to this task, and the most relevant (i.e., discriminating) attribute is the SIZE (both length and width).

9.2.4 Sources of Information

Six sources of information have been selected to cover the diversity of commonly available maritime surveillance sources, ranging from hard (physical sensors) to soft (humans) sources, providing either subjective or objective information, either in a numerical or qualitative (using natural language) format.

[2] Basing the relevance assessment on the attributes only does not cover the information content. Indeed, a piece of information of very low informational content is not relevant as it would barely impact the previous belief state.

Moreover, sources are a combination of an information container and an information provider [18], which allows distinguishing between automatic processors (such as trackers or classifiers) and human analysts processing the same initial signal or image. Sources are further characterised according to their expertise defined by the list of attributes about which they are able to provide information [18], among the list of the five basic attributes. For each track, each attribute is reported by either two or three distinct sources, providing the player with both some redundancy and complementarity in information received. Table 9-1 lists the six sources of information, their expertise (in terms of the attributes about which they are able to provide information) and their range in terms of the track covered (VA, VB or both). No specific information about the sources' quality (in terms of reliability) is provided but the player is expected to rely on prior own knowledge (or perception) about this aspect and process the information accordingly.

Table 9-1: Sources Coverage and Expertise.

Sources		**Range**	**Expertise**				
Container	*Provider*		LOCATION	HEADING	SPEED	SIZE	TYPE
Radar	Operator C	*VA* and *VB*		X		X	
	Tracker/ATR		X		X		X
SAR	Analyst A		X	X			
	ATR				X		
Camera	Analyst B	*VA* only	X	X	X	X	X
Cargo	Captain	*VB* only	X	X	X	X	X

9.2.5 Information Quality Dimensions

The comprehensive consideration of all Information Quality Dimensions (IQDs) [19], [20], [21], is beyond the purpose of the Risk Game, and we focus on the three dimensions below assumed to have a possible impact on the reasoning and decision processes and that we consider as basic and independent:

1) **IQD1: Trueness/Falseness** – Refers to the "true" (or reference) value of a vessel attribute and is a synonym for "correctness".
2) **IQD2: Certainty/Uncertainty** – Refers to a degree of confidence or belief assigned to a specific value (or set of values) to be "true". Its cause can be either a lack of knowledge (epistemic uncertainty) or the random variability of the underlying process (aleatory uncertainty). When assigned by the source itself it may be called "self-confidence".
3) **IQD3: Precision/Imprecision** – Refers to a set of possible values: The smaller the cardinality of the set (or the length of the interval), the higher the precision. It reflects the inability of the source to provide a single value or to discriminate between several values and is a synonym for "non-specificity".

On the one hand, imprecision (or precision) and uncertainty are opposed [20]: "I'm certain that the speed of the vessel is between 3 and 6 knots" (imprecise but certain statement) vs. "I'm not certain that the speed of the vessel is 5 knots" (precise but uncertain statement). On the other hand, precision and trueness are often associated in performance assessments of systems, gathered under the notion of accuracy[3]. These three IQDs will be our basic independent variables of the Risk Game.

[3] When referring to a series of independent tests, accuracy refers to a combination of trueness and precision in ISO 5725 [22].

In addition to the three above-selected IQDs, the following four dimensions will also be considered, assuming they could also have an impact on the reasoning and decision processes. They are latent variables as they result from some variations of the observed (and controlled) IDQ 1 to IDQ 3:

1) **IQD4: Conflict** – Relative to some inconsistency between the different pieces of information. Indeed, because some pieces of information may be false, they may conflict with others.
2) **IQD5: Relevance** – Relative to the output of the reasoning process in general. It can be relative to the goal (here the decision to be made), or to the belief state of the player. Understood as informational content, relevance directly arises from **imprecision** and **uncertainty**. Relevant information impacts previous belief, or is helpful to make the decision.
3) **IQD6: Reliability** – Relative to how much a source's statement can be relied on, which could be derived from the ability of the source to provide correct information (previously assessed or known).
4) **IQD7: Independence** – Relative to the correlation of the sources (i.e., if the output of one source influences the output of another one), and relative to the correlation of attributes (e.g., TYPE and SIZE are dependent attributes).

The relevance of the input pieces of information will not explicitly vary, but will vary implicitly with the type of attribute and the uncertainty or imprecision of the piece of information. For instance, information about the location is less relevant to identify the vessel than information about its type. Similarly, the sources' reliability will not be explicitly specified but due to their diversity, the sources exhibit different reliability degrees: For instance, the camera analyst may be considered more reliable than the cargo captain, or the radar providing the location may be considered more reliable than the analyst of the SAR imagery.

Moreover, some of the attributes reported are not independent from each other. In particular, the instant components of LOCATION and HEADING can reasonably be considered independent from each other and from the other attributes while SPEED, SIZE and TYPE are dependent. Figure 9-3 displays the different variable of the Risk Game and their dependencies.

Note that a suitable formalisation of the reasoning process supported by a dedicated mathematical framework to reasoning under uncertainty would allow the measurement of these quality dimensions. This is however out of the scope of the current chapter and will be addressed in future work.

9.2.6 Levels of Information Quality

For the sake of simplicity (i.e., to keep the number of combined quality levels tractable), we consider only two possible states for the three basic IQD variables: Pieces of information are either perfectly true, precise, certain, or moderately false, imprecise, uncertain, leading to a set of eight possible combinations:

- **Falseness** is a shift from the true value but is always in the range of acceptable mistakes so that it is never obvious that the PoI is actually false. For instance, no source reports that the observed vessel is of length 150 meters or that it is a Ferry.
- **Uncertainty** scale follows the Standardised Lexicon used by the US National Intelligence Council, and we consider only two values, likely and almost certain, to which we assign the two corresponding numerical degrees of 0.6 and 1: Soft sources (humans) report phrases and hard sources (sensors or algorithms) report numerical values for an equivalent meaning. For instance, "likely" exactly means "confidence of 0.6".
- **Imprecision** is represented either by intervals for the numerical attributes (SPEED, LENGTH, WIDTH) or sets of values for nominal and discrete attributes (LOCATION, HEADING, TYPE). Moreover, some soft sources report fuzzy statements (Small, Medium, Large for the SIZE, for instance).

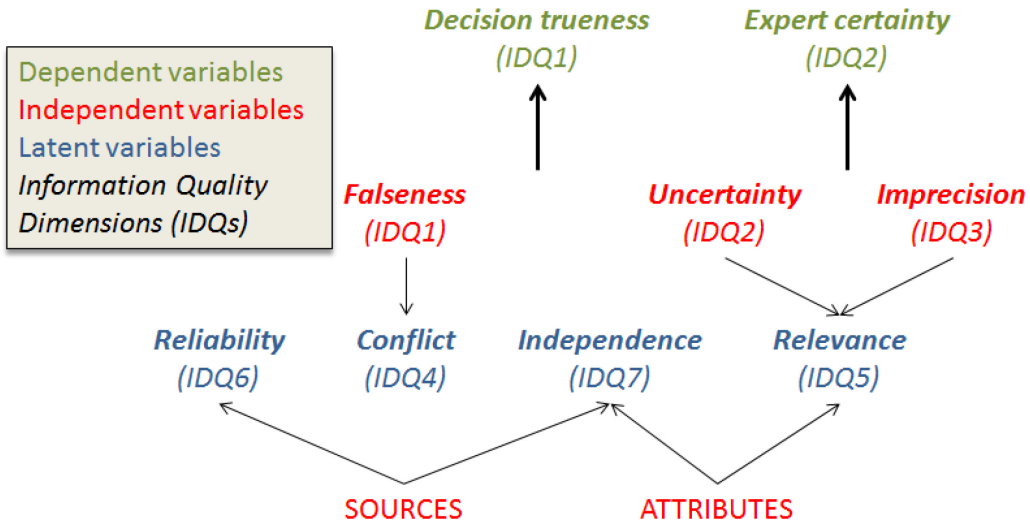


Figure 9-3: Independent, Latent and Dependent Variables of the Risk Game.

Note: Nine quality dimensions (italic) are considered: three are independent variables (red, italic) controlled randomly through dice roll; four are latent (blue, italic), derived through the other two independent variables (Sources and Attributes) controlled by the players' queries, and two are the dependent variables (green, italic) which are measured during the experiment.

The information quality the player obtains is randomized over the eight versions of the same piece of information through a dice roll, as displayed in Table 9-2. The best and worst quality levels are assigned a lower probability. Also, a higher probability is assigned to the true pieces of information either only imprecise or only uncertain. This probability distribution is certainly a bit pessimistic regarding sources' performance in general. However, the global low quality of information is explained by the rough sea state as well as the high distance between the sources and the vessels relative to their range.

Table 9-2: Eight Information Quality Levels and Corresponding Randomization.

True	Precise	Certain	Randomization
1	1	1	0.11
1	1	0	0.22
1	0	1	0.22
1	0	0	0.11
0	1	1	0.06
0	1	0	0.11
0	0	1	0.11
0	0	0	0.06

9.2.7 Information Cards and Fusion

A Piece of Information (PoI) is denoted by *f* and is represented by a tuple:

$$f = \langle s, v, a, x, p, c, t \rangle \tag{9-3}$$

where *s* is the source providing the piece of information, *v* is the vessel (or track), *a* is the attribute, *x* is the estimation or measurement of *a* for *v* by *s*, *p* is the precision of *f*, *c* is its certainty and *t* is its trueness. For instance,

$$f = \langle (Radar, Tracker), VA, SPEED, 4knots, 1{,}0{,}1 \rangle \tag{9-4}$$

reads: "The radar with its associated tracker estimates the speed of Vessel *A* at 4 knots with a confidence of 0.6", and this information is true. The last element of trueness is obviously not provided to the player and known only to the game facilitator.

Pieces of information are abstracted through *information cards* available for query to the player. Each card is uniquely identified by a coded ID (Figure 9-4).



Figure 9-4: An Example of an Information Card Abstracting a Piece of Information About the Location of Vessel *A* Provided by the Tracker Processing the Radar Signal.

First, only the back of the card (black) is visible to the player. Then, the player selects:

1) The track to query;
2) The attribute about which to obtain information; and
3) The source providing information.

Thus, at each round, the player has 3 degrees of freedom to select the piece of information (i.e., the card): The track (or vessel), the attribute and the source. These point towards a single pile of eight cards from which information quality varies. After rolling two dice, the player is given the card of corresponding quality by the game facilitator and can read the information content (white part). The player is now ready to rate the belief degree towards events *A* and *B*.

The information processing to be performed by the player to reach suitable situation awareness can be formulated as fusion task:

$$f_P = \oplus_n(f_n) \tag{9-5}$$

where f_P would represent information aggregated by the player P from the individual pieces of information f_n queried and $\oplus$ denotes some aggregation operation used by the player to derive f_P from the set of f_ns. f_n is the piece of information under the form of Equation (9-4) f_P can be further synthetised as a pair (Bel(A); Bel(B)) of belief degrees towards events A and B respectively.

9.2.8 Gathering of Belief States and Decision Making

In practice, the Officer of the Watch would have all this information (i.e., the five attributes values for the two vessels) available at the same time, possibly displayed on several screens. In the Risk Game we would like to decompose the reasoning process to the granularity level of individual pieces of information, in order to track the successive belief states of the player while discovering and processing the different pieces of information. It is thus explained to the player that time is fixed, the situation does not evolve, and the time taken for processing information does not matter.

Once the player is given the information card by the facilitator and has read the message, the belief can be assessed about events A and B through the form displayed in Figure 9-5. A scale between 0 and 1 with steps of 0.2 is proposed to the player. The player is expected to rate both events at each time round, regardless the vessel queried, but is not reminded to do so which may result in missing data. One row corresponds to one game round: the first column corresponds to the ID of the card which captures the track, attribute, source and the information quality (independent variables), the red section captures the player's belief degree about event A and the blue part about event B. A cross indicates the player's belief state, the dependent variable. The player can select as many cards desired (up to the 26 cards available) and when ready to decide, can check the corresponding box at the bottom of the form and the game is over. The set of forms filled by the players and collected during the Risk Game constitutes the dataset to be analysed in Section 9.4.

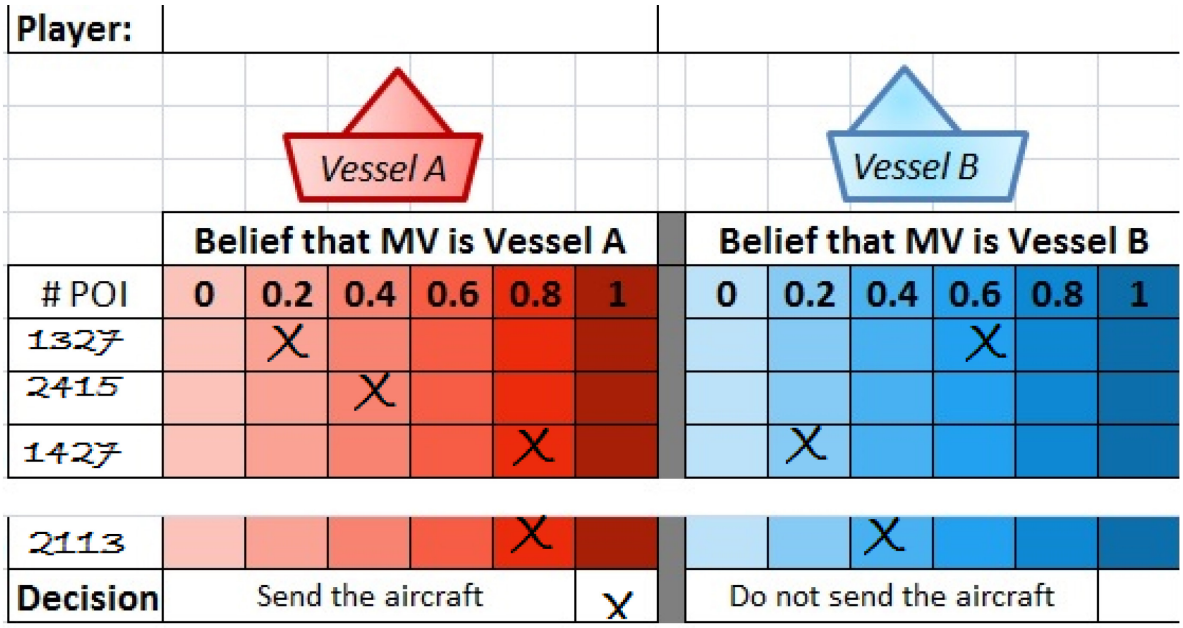


Figure 9-5: Belief Assessment Form to be Filled in by the Player After Discovering Each Piece of Information. "MV" stands for "Missing Vessel" (or lost vessel).

9.3 METHODOLOGY ASSESSMENT

A rehearsal session with some experts allowed validating the scenario, the set of sources as well as the set of pieces of information, making the game quite realistic. The players of the rehearsal were generally happy with the "draft" design of the game, quickly caught the rules, and validated the story and sources reports. They were committed at the beginning of the game and quite excited during the play.

A total of 32 players (different from the rehearsal) selected as experts in maritime surveillance and threat assessment played the Risk Game during a Tabletop Exercise (TTX) held at the NATO STO Centre for Maritime Research and Experimentation (CMRE). A survey was conducted in which the players were asked to assess the Risk Game according to the criteria of *Realism*, *Operational relevance*, *Understandability*, *Engaging ability*, *Elicitation efficiency* and *Facilitation*. Figure 9-6 displays the players' feedback, the darker the purple colour the better the feedback. In the light of this survey, the players generally found the Risk Game either *extremely realistic* or *very realistic* (60%) and either *extremely* or *very operationally relevant* (80%). They all found the elicitation method *effective* while 72% of them found it either *extremely* or *very effective*. They also all understood the purpose of the game only 20% moderately understood it. Most of them (96%) found the game at least *moderately engaging*, while 88% found it either *extremely* (44%) or *very engaging* (44%). Finally, they generally appreciated the facilitation (92% found the game either *extremely* or *very well facilitated*).

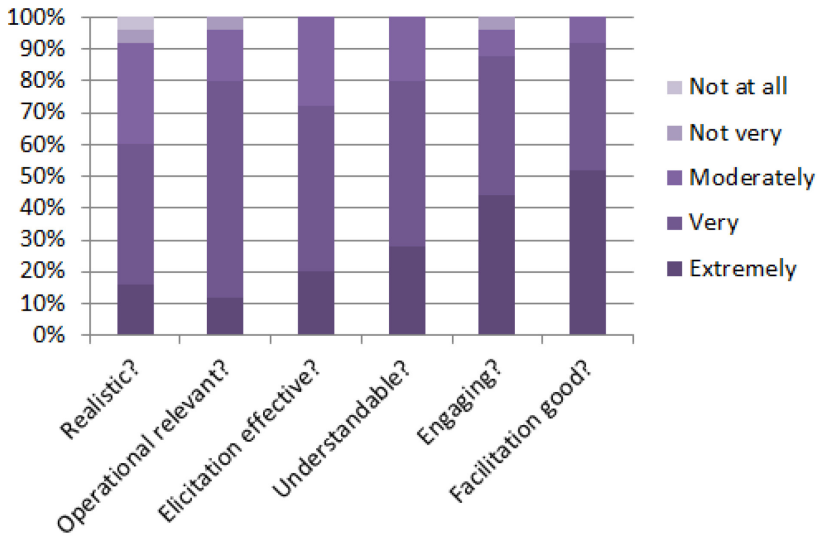


Figure 9-6: Players' Feedback After the Risk Game.

9.4 EXPLORATORY DATA ANALYSIS

We provide below some results from the exploratory analysis of the data collected during the game. Due to the high number of variables of the game for information queries, information quality and decision quality, we provide examples of the kind of analyses that can be derived from the Risk Game, while an exhaustive analysis of the data is left for future work. The results and analyses presented here are thus a subset only of the interesting and relevant ones. They nevertheless cover the main aspects of information quality and its impact on decision making.

We first investigate the queries from the players which illustrate their information needs (Section 9.4.1), followed by the final belief states and decision made (Section 9.4.2) and finally information quality (Section 9.4.3).

9.4.1 Analysis of Queries

The players had four degrees of freedom to query the information: The track (or vessel) *A* or *B*, the attribute (LOCATION, HEADING, SPEED, TYPE and SIZE), the source (Radar, SAR, Camera and Cargo), and the number of queries. Also, they had a complete freedom in the order of their queries. The analysis of the players' query behaviour is meaningful to highlight their information needs, their perception (or *a priori* knowledge) of sources' quality and attribute relevance to the problem, as well as their reasoning strategy.

9.4.1.1 Information Needs

Each player selected a variable number of cards as displayed by the histogram of Figure 9-7. The average number of PoIs is 13, with a standard deviation of six queries. There is considerable variability in the number of queries, some players selecting a very few PoIs while other selected all available ones. If we compute artificially the time of reaction to the number of PoIs retrieved and processed, it is interesting to note that the "quickest" and the "slowest" players made the same decision that is to send the patrol aircraft, by querying 3 and 26 PoIs respectively.

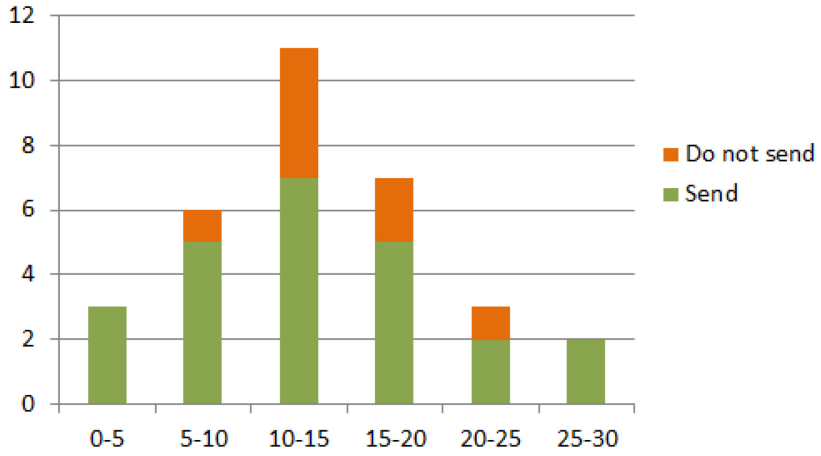


Figure 9-7: Number of Queried Pieces of Information Based on the Final Decision Made.

The decision to send the patrol dominates the entire range of numbers of PoIs. Table 9-3 shows the ratio of the two vessels queried by the players, at their first query (left column) and the most queried (right column).

Table 9-3: Ratio of Vessels Queried by the Players.

	First	Most
Vessel *A*	72%	84%
Vessel *B*	28%	12%

Unsurprisingly, Vessel *A* appears to be the priority vessel to the players (first queried for 72% of the players, and most queried for 84%), as indeed it is the one in the area of responsibility.

The perceived relevance of information perceived by the players can be estimated through the analysis of the attribute queries. The ratio of the first and most queried attribute by the players regardless of the vessel is displayed in Table 9-4. The number of queries has been normalised to the number of available PoIs, either 3 or 2 (see Table 9-1).

Table 9-4: Ratio of Attributes Queried by the Players.

	First	Most
LOCATION	81%	53%
SPEED	3%	6%
HEADING	13%	6%
TYPE	3%	28%
SIZE	0%	6%

We observe that the location was perceived as the most relevant attribute by the players (both first and most queried), although it was the least relevant to solve the identification problem (see Section 9.2.2). Even if the players had initially a rough idea of the location of each vessel, it seems that a more precise location was needed, maybe to better visualise the scene, to better appreciate the closeness of the vessel to the coast. We do not have so far data to support any of these hypotheses but only the feedback of the players after the game to help clarify this behaviour. Table 9-5 displays the ratio of queried PoIs of the two categories of attributes: The category "behavioural" corresponds to LOCATION, HEADING and SPEED, which help in clarifying the vessel behaviour, while the category "classification" includes SPEED, SIZE and TYPE.

Table 9-5: Ratio of Attribute Categories Queried by the Players.

	Most
Behavioural	60%
Classification	40%

This table relates the effective relevance of the attributes (as defined in Section 9.2.2) and the perceived relevance if we consider that the players would query more frequently the information perceived as relevant. Most queries focused on some attributes relevant to anomaly detection and threat assessment in general (i.e., behavioural attributes) but less relevant to the purpose of the game, which was rather a classification problem (i.e., the classification attributes would have been more helpful to discriminate between the two vessels).

We assume that the expected reliability of sources can be estimated by the first and most queried source among Radar, SAR, Camera and Cargo captain, as displayed in Table 9-6. These results provide an idea of the trust or confidence of the players in the different sources. The radar appears as the most trusted source (first and most queried). The cargo captain and the camera analyst received significant interest as well, as it appears that human information was very valuable to the players.

Table 9-6: Ratio of Sources Queried by the Players.

	First	Most
Radar	63%	72%
SAR	13%	22%
Camera or Captain	26%	6%

However, these results must be analysed carefully since the source query is not independent from the attribute query (i.e., not all the sources provided information about all the attributes). Indeed, the study could be refined to consider couples (source, attribute) as it became clear during the game that the players rely on a source to provide a given attribute and on another for another one.

9.4.1.2 Query Strategies

Based on the way the players queried the information, we observed different reasoning strategies (without being able to judge about their respective quality, though). As an example, Figure 9-8 shows the ratio of switches in the queries between Vessels *A* and *B*, either from *A* to *B* or from *B* to *A*, relatively to the maximum number of possible switches. A null ratio means that the player queried a single vessel (either *A* or *B*), thus never switched. A low ratio means that the player mostly queried one vessel and then the other one. A high ratio means that the player systematically queried one vessel and right after the other one, demonstrating a reasoning strategy by comparison. This later group of players probably took advantage of the evidences against one event to increase their belief for the other one. It is still very difficult however to confirm this hypothesis in the light of the data gathered.

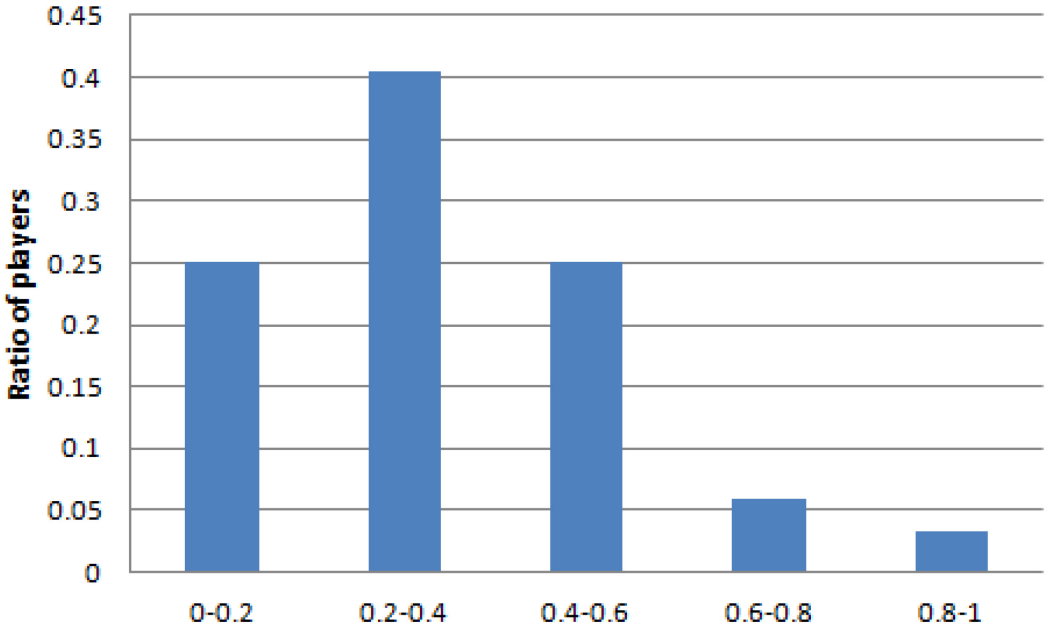


Figure 9-8: Ratio of Switches in Queried Information Between Vessel *A* and Vessel *B*.

Because all attributes are available from either two or three sources, the players have some control on the degree of *redundancy* and *complementarity* of the information gathered. An empirical way to estimate the need in redundant information is to count the number of queries toward the same attribute from different sources (for a given vessel). Analogously, an empirical way to measure the level of the complementarity in information was derived by computing the number queries on different attributes. These two dimensions of redundancy and complementarity are jointly plotted in Figure 9-9. Each dot represents a player, together with the final decision made (blue cross or red bullet). The need for redundancy however may be explained by a "bad" piece of information, conflicting with previous ones, which requires further investigation. On the other hand, complementarity is needed to cover the diversity of attributes and that is what most of the players were looking for. We observe that most of the players adopted a strategy based on a high level of complementarity (i.e., [0.8 – 1]) and a medium level of redundancy, giving priority to multi-attribute investigation.

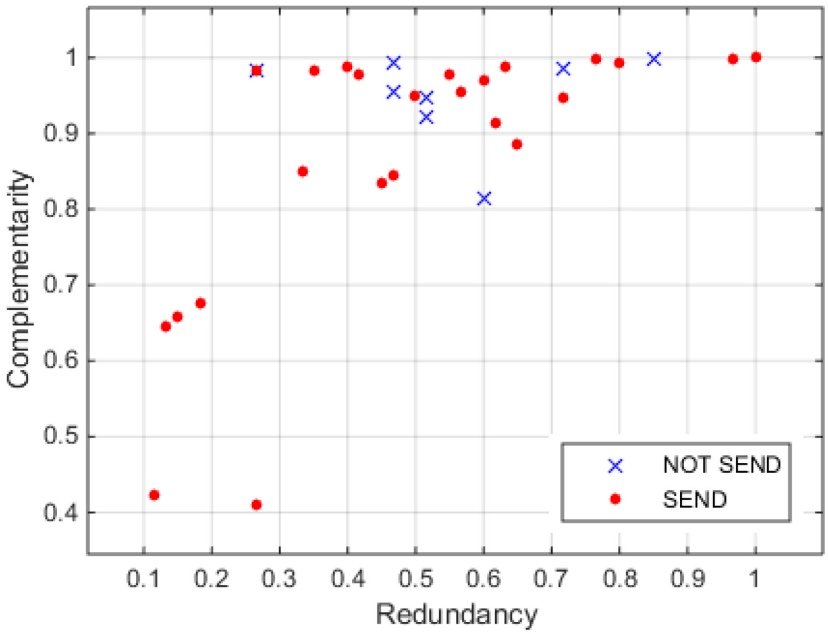
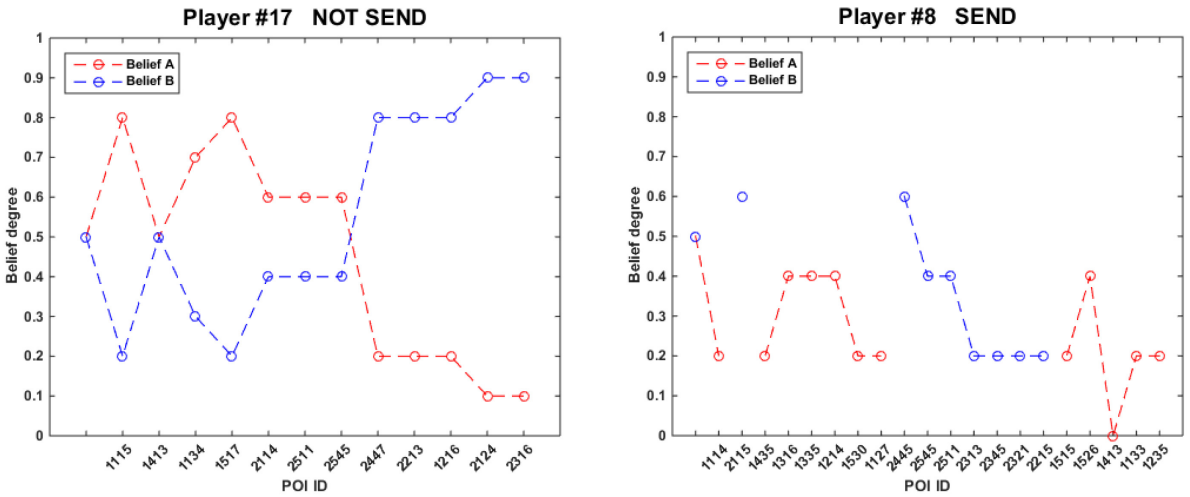


Figure 9-9: The Redundancy/Complementarity Values Grouped, Based on the Decision Made.

9.4.2 Belief States and Decision

Two kinds of output were recorded from the players: (1) Their sequential belief states relative to both events *A* and *B* and assessed after each step *n* of information discovery and (2) the decision to send or not to send the patrol aircraft once they felt they knew enough about the situation. Figure 9-10 provides a sample of the data recorded for two players. Player #17 (Figure 9-10(a)) systematically assessed both events (no missing assessment), symmetrically rated events *A* and *B*, and while initially generally leaned toward event *A* (with a step of high uncertainty), finally changed mind to reach a high belief degree toward *B*. Player #8 (Figure 9-10(b)) generally assessed only one of the two events, went through a phase of high confidence for event *B* to finally reached a state of high uncertainty.



(a) Player #17 Belief Assessment. **(b) Player #8 Belief Assessment.**

Figure 9-10: Two Examples of Sequential Belief Assessments for Both Events *A* (Red) and *B* (Blue) for Two Players. The ID of the card is displayed on the x-axis.

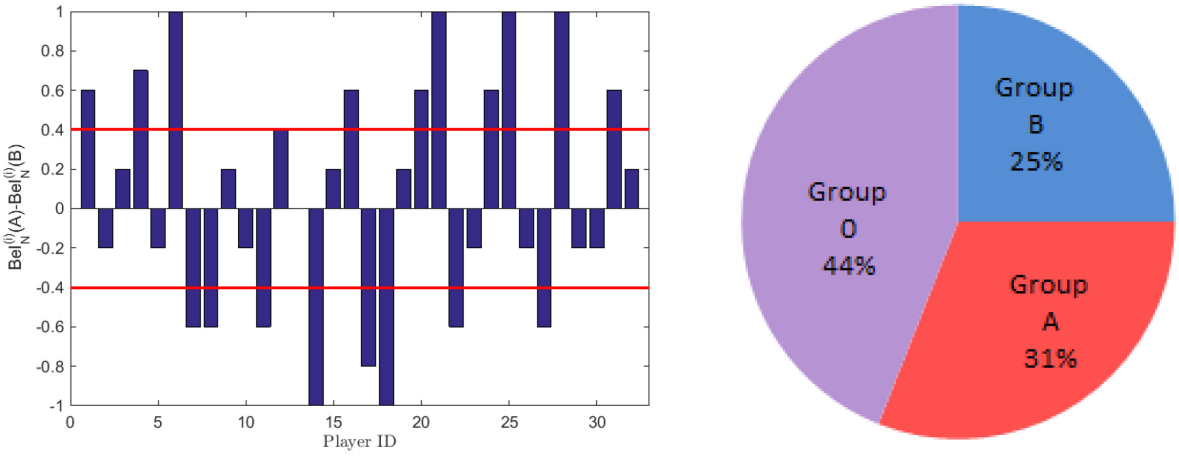
Because of the binary problem (either *A* or *B*), we will consider only trueness and certainty (ignoring imprecision possibly expressed by the player). Trueness is simply estimated relative to the known correct track, while certainty is assessed through the last belief state before decision. If we denote by $Bel_n^{(i)}(A)$ and $Bel_n^{(i)}(B)$ the belief degrees of Player i at step n, for n varying from 1 to N, toward events A and B respectively, assuming these belief degrees range from 0 to 1, the certainty degree of the player will be defined by:

$$c^{(i)}(n) = |Bel_n^{(i)}(A) - Bel_n^{(i)}(B)| \tag{9-6}$$

so that $c^{(i)}(n) = 1$ means that the player is totally certain about one event or the other, and $c^{(i)}(n) = 0$ means that the player is totally uncertain (does not lean toward an event or the other). Figure 9-11(a) shows the final belief state (at step *N*) of each player. We partitioned then the set of players according to their final belief leading to three groups (see Figure 9-11(b)):

- **Group B**, denoted as G_B: Strong belief for event *B*, for $Bel_N^{(i)}(A) - Bel_N^{(i)}(B) \leq -0.4$. A negative value means that the player has a stronger belief for event *B* before deciding, i.e., low uncertainty toward *B*.
- **Group A**, denoted as G_A: Strong belief for event *A*, for $Bel_N^{(i)}(A) - Bel_N^{(i)}(B) > 0.4$. A positive value means that the player has a stronger belief for event *A* before deciding, i.e., low uncertainty toward *A*.
- **Group 0**, denoted as G_0: High uncertainty regarding *A* and *B*, for $-0.4 < Bel_N^{(i)}(A) - Bel_N^{(i)}(B) \leq 0.4$. A null value means that the player has equal belief for both *A* and *B*, i.e., a high uncertainty.

In our scenario, Vessel *A* was indeed the missing vessel and thus should have been assessed as a threat. Vessel *B* was another fishing vessel from Right Land of the same type but slightly smaller than the missing vessel, and was going back to its port.



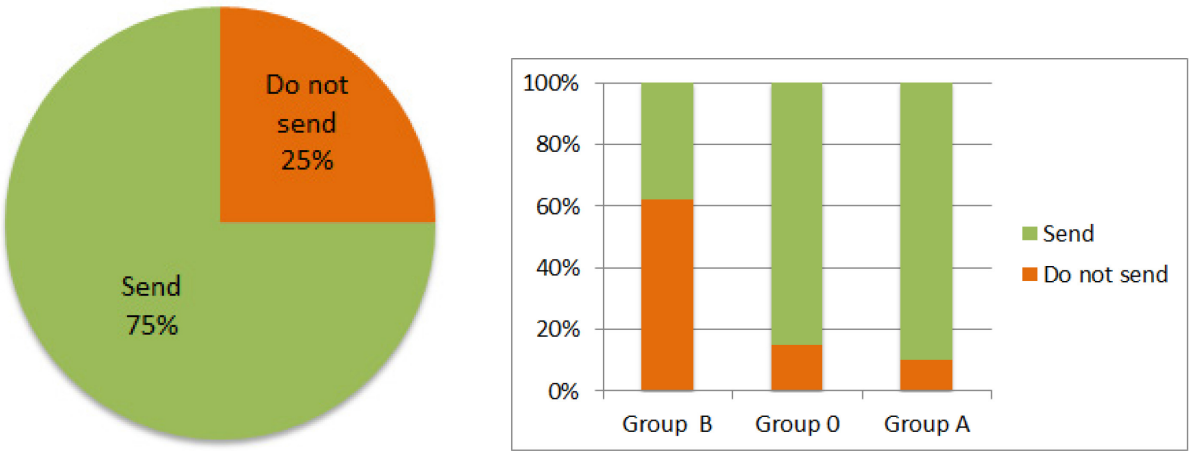
(a) Final Belief State for Each of the Players.

(b) Players Groups Based on Final Belief State: GB, G0 and GA

Figure 9-11: Final Belief State Before Decision.

Figure 9-12 provides a summary of the decision made by the players (green for decision "Send" and orange for decision "Do not send"). Figure 9-12(a) shows that the majority (3/4) decided to send the patrol aircraft. This result is not surprising due to the asymmetry in the two vessels' risk level: Only Vessel *A* was possibly

at risk since Vessel *B* was far and not in the AOR. Figure 9-12(b) refines the decision made within the three groups defined above. We first observe that while we would expect that a strong belief toward *A* (group G_A) would lead automatically the players to decide to send the patrol a small percentage actually did not. When asked about a justification for such a decision, some players answered that they still had time to send the patrol and delayed their decision at that time due to its cost. We can make a symmetrical observation about players who strongly believed that the missing vessel was still in its area (event *B*) but who actually decided to send the patrol. Their decision was justified by the contextually risky environment which implied a cautious decision to avoid missing any suspicious event.



(a) Decision Made by the Players

(b) Decision Relative to Three Groups of Final Belief States

Figure 9-12: Decision and Relation to Final Belief State.

Finally, we observe that of the players with high uncertainty before decision (group G_0) a large majority decided to send the patrol aircraft. This highlights the strong influence of the context as noticed at the beginning of this analysis section: Because the Harbour Protection Level is still TWO, the decision favoured was to send the patrol.

9.4.3 The Effect of Information Quality

We analysed the effect of information quality on both the decision and final belief state with two different approaches. First, we performed an analysis player by player considering the final belief state only. Then, we analysed each piece of information individually regarding its impact on previous beliefs, regardless the player.

9.4.3.1 Falseness

In Figure 9-13 the impact of the ratio of false information (over the number of pieces of information picked up) on the final belief state: The darker, the higher the ratio of false information. We observe that the players who leaned toward event *B* (group *B*) indeed received a higher ratio of false pieces of information than the ones who were highly uncertain (or confused, group 0), or than the ones who leaned toward event *A* (group *A*). This result confirms our expectations that a high ratio of false pieces of information increases the confusion in the decision-maker mind (as group 0 players were more uncertain than groups *A* and *B*), up to "mis-assessing" the situation (as group *B* players were rather certain of the wrong hypothesis).

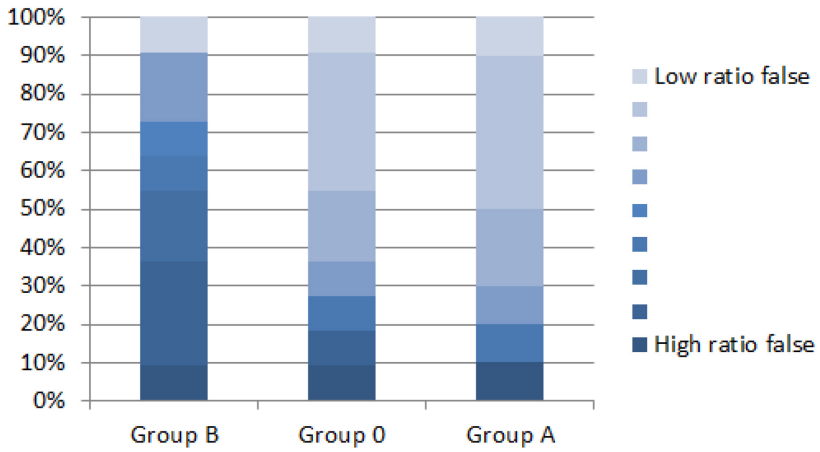


Figure 9-13: Impact of False Information Ratio on the Final Belief State.

9.4.3.2 Certainty and Precision

Figure 9-14 displays the impact of information content only (i.e., regardless of the falseness of the information) on the belief change in the player's mind it produced (event *A* only is shown). We considered three categories of belief change:

- **Null belief change** when the belief about *A* did not change at all, i.e., $|Bel_N^{(i)}(A) - Bel_{n-1}(A)| = 0$, where $Bel_n^{(i)}(A)$ is the belief if the player *i* at the step *n* about Vessel *A* and $Bel_n^{(i)}(A)$ is the belief if the player *i* at the previous step *n*-1 about Vessel *A*.
- **Low/Medium belief change** when the belief about *A* changed but remained in favour of the same event, i.e., $0 < |Bel_N^{(i)}(A) - Bel_{n-1}(A)| < 0.5$
- **High belief change** when the belief about A changed so that the player changed his mind either from *A* to *B* or from *B* to *A*, i.e., $|Bel_N^{(i)}(A) - Bel_{n-1}(A)| \geq 0.5$

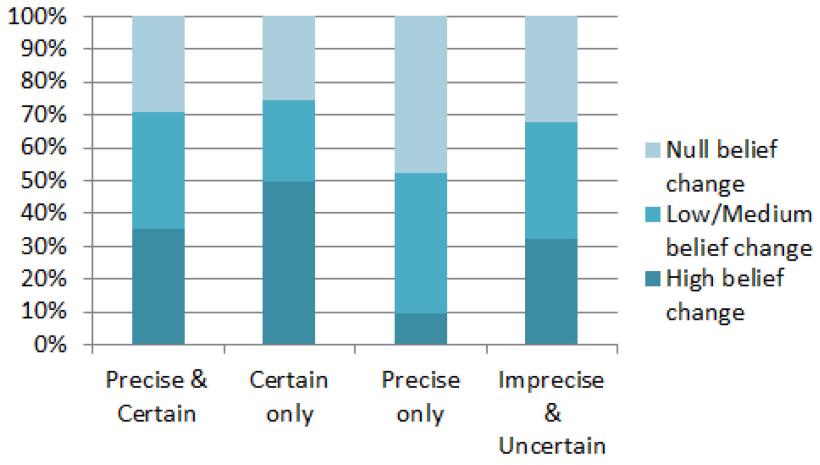


Figure 9-14: Impact of Information Content Only on the Belief Change (About Event *A*).

The relative ratio of pieces of information which produced either no belief change or a low/medium belief change respects more or less the available information quality (see Table 9-2). More interesting however is the group of pieces of information which made the players switching from a belief toward *A* to a belief

toward *B* (or reversely), darker blue in Figure 9-14. It appears that this group contains a very high proportion of *certain* pieces of information. In other words, the set of pieces of information that were certain only (i.e., not imprecise), represented by the second column in the figure, were more likely to make the players change their belief from one event to the other one. Although further investigations are required, it seems that information expressed with high certainty has a high impact on the players' belief change.

9.4.3.3 Relevance

We interpret a frequent query of a specific attribute as having some level of *perceived relevance* by the player to the problem. We highlighted that although the behavioural attributes were generally frequently queried, they did not contain many relevant discriminant elements which were rather carried by the classification attributes (TYPE and SIZE) since the two vessels differed mainly in their length and width.

We thus analysed the impact of attributes on the belief change and especially investigated which attributes provoked the highest belief change such that the players changed their mind from one event to another (either *A* to *B* or *B* to *A*). Figure 9-15 displays the ratio of attributes which indeed made the players change their minds. It is interesting to observe that unsurprisingly the size has the most impact followed by the type. What is surprising though is that the location had a similar impact to the size while it in fact did not contain any information, which would have helped the player discovering where the missing vessel was.

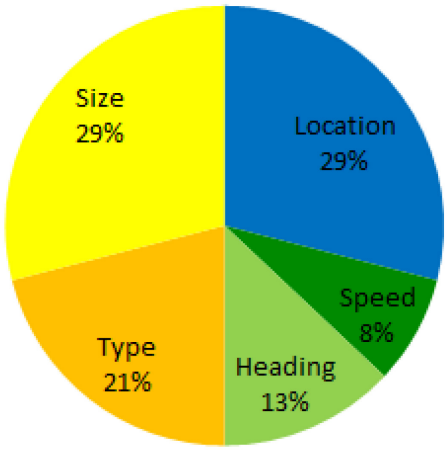


Figure 9-15: Impact of Attribute on Belief Change from *A* to *B* or *B* to *A*.

9.4.4 Possible Biases

The results presented here should be interpreted in the light of possible biases induced by the game design. In particular:

- **Operational context**: The global Harbour Protection Level (TWO) strongly influenced both the belief assessment and decision toward event *A* and "Send the patrol" respectively. Further study would change the operational context only (in the case of Search and Rescue, for instance) where the events are symmetric in terms of risk assessment, in such a way that the decision and belief assessment would be guided only by the information at hand and the concern of a timely and accurate intervention.
- **Belief assessment**: A few players might have misunderstood the rules or the form to be filled and possibly confounded the instant belief assessment with the assessment based on accumulated pieces of information. This may be an important bias that we could have observed for one player whose belief assessments did not exhibit an evolution.

- **Complexity**: The number of information quality dimensions is probably too high, so it became tedious to find interesting results in the data gathered. These became obvious after partitioning the data. We could expect that a simpler design, with a reduced number of variables would simplify the analysis and lead to more direct link exhibiting the impact of information quality of human belief assessment.
- **Limited number of players**: The board version of the game has the advantage of a close interaction with the experts who additional provide explanations of their reasoning path and justification for their decision. This aspect of the facilitation is extremely beneficial to understand the operational concerns of the players. However, the small number of samples gathered limits strongly the scope of the conclusions drawn. A digital version of the game would allow a deeper study and more convincing results supported by meaningful statistics.

9.5 CONCLUSION

We presented the Risk Game, a methodology to elicit experts' knowledge and know-how in threat assessment and decision making in risky environments, with imperfect information. The Risk Game was designed to highlight in particular the impact of information quality on belief and threat assessment and decision making. We selected the three quality dimensions of *falseness*, *uncertainty* and *imprecision* that we made vary independently. The other dimensions of *relevance*, *source reliability* and *conflict* were also studied as naturally derived from the three basic dimensions.

Thanks to the involvement of the 32 players, we have been able to validate:

1) The Risk Game as an entertaining elicitation method;
2) The efficiency of the approach to gather data demonstrating the impact of different aspects of information quality on belief assessment and decision making; and
3) The framework for reasoning analysis establishing formal links between different information and source quality dimensions, as well as between context, threat assessment and decision.

In the light of the data analysed, we have been able to observe some interesting aspects. First, the analysis of the players queries through information cards about source, attribute and vessel dimensions demonstrates the distinction between *information needs* (which information the players think is useful to solve the problem) and the *effective relevance* of information (which information actually helps solve the problem). It appears that while positional information (location, heading and speed to some extent) was less useful than the classification information (size and type), it was the most queried by the players. We also highlighted the information quality dimensions of *complementarity* and *redundancy*, as a natural basis of the human fusion process. It was shown that most of the players advocated for a medium redundancy by high complementarity. From the queries analysis, we set up the basis to underline different reasoning strategies, i.e., *comparative reasoning* vs. *single event focus*.

Second, we analysed the impact of the information quality along the three dimensions of falseness, uncertainty and imprecision, on the uncertainty state of the player before decision and upon making the decision. The decision to send the patrol was made by 75% of the players, which actually was the correct decision given our scenario. We explained this asymmetry by the context (i.e., Harbour Protection Level of TWO) which made the players favour this safer decision in case of high uncertainty, as supported by the results. We also clearly observed that the higher amount of false information, the higher the uncertainty before decision.

We finally characterised the impact of the individual pieces of information quality on the belief change they induced in the player's mind. It appeared that information expressed with *certainty* by the sources had a high impact on the players' instantaneous belief.

Also, we confirmed that the most effective relevant attributes (size and type) had very high impact in belief change, together with the location. This last result is surprising since the location cards did not contain any effectively relevant information which would have help the player discriminate between the two vessels. This maybe stresses the need for humans to visualise the scene to better grasp the situation. Certainly, further investigation would be required to be able to draw any valid conclusion.

In future work, we will develop methodology to focus on other specific aspects of information quality (e.g., source quality, inconsistency). We will analyse the reasoning profiles of the players by comparing their assessment with automatic reasoners. Finally, a formalisation of the process will lead to deeper and quantitative analysis of information quality on belief assessment.

9.6 ACKNOWLEDGEMENTS

The authors would like to thank the Allied Command Transform (ACT) for funding this work. Appreciation is given to the Italian Navy under the direction of Lt. Cdr. Claudio Cuomo (COMFORDRAG) for advice and feedback given during game design, rehearsal and execution. Their suggestions helped us to provide an improved version of the game. To Cdr. Mike Ilteris (US) and Lt. Cdr. Nick Gwatkin (UK) whose operational experience helped to validate the information format and content. Appreciation is also given to the CMRE scientific staff who contributed to the smooth running of the Tabletop Exercise (TTX) in support to Harbour Protection in 2014. Finally, appreciation is given to all players for their enthusiasm and commitment to a new and experimental game format and for sharing their invaluable knowledge and experience throughout the game-play.

9.7 REFERENCES

[1] Jousselme, A.-L., Pallotta, G., and Locke, J. (2015). A risk game to study the impact of information quality on human threat assessment and decision making. Technical Report CMRE-FR-2015-009, NATO STO Centre for Maritime Research and Experimentation, La Spezia, Italy.

[2] Steinberg, A.N., and Bowman, C.L. (2001). Revisions to the JDL data fusion model. In: *Handbook of Multisensor Data Fusion*, The Electrical Engineering and Applied Signal Processing Series, Hall, D.L., and Llinas, J. (Eds.), 2-1 to 2-19. Boca Raton, FL: CRC Press.

[3] Borek, A., Parlikad, A.K., Woodall, P., and Tomasella, M. (2014). A risk-based model for quantifying the impact of information quality. *Computers in Industry,* 65(2):354-366.

[4] Raghunathan, S. (1999). Impact of information quality and decision-maker quality on decision quality: A theoretical model and simulation analysis. *Decision Support Systems,* 26(4):275-286.

[5] Moges, H.-T., Dejaeger, K., Lemahieu, W., and Baesens, B. (2013). A multidimensional analysis of data quality for credit risk management: New insights and challenges. *Information & Management*, 50(1):43-58.

[6] Ceruti, M., Das, S., Ashenfelter, A., Raven, G., Brooks, R., Sudit, M., Chen, G., and Wright, E. (2006). Pedigree information for enhanced situation and threat assessment. In: *Proceedings of the 9th International Conference on Information Fusion*, Florence, Italy.

[7] Wang, R.Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, 41(2):58-65.

[8] Carballo-Cardenas, E.C., Mol, A.P., and Tobi, H. (2013). Information systems for marine protected areas: How do users interpret desirable data attributes? *Environmental Modelling & Software,*41:185-198.

[9] Hoffman, R.R., Shadbolt, N.R., Burton, A.M., and Klein, G. (1995). Eliciting knowledge for experts: A methodological analysis. *Organizational Behavior and Human Decision Processes,* 62(2):129-158.

[10] Cao, Y. (2014). Ludic elicitation: Using games for knowledge elicitation. Ph.D. thesis, State College, PA: The Pennsylvania State University.

[11] Schönbohm, A., and Jülich, A. (2016). On the effectiveness of gamified risk management workshops: Evidence from German SMEs. International *Journal of Serious Games* 3 (2).

[12] Taillandier, F., and Adam, C. (2018). Games ready to use: A serious game for teaching natural risk management. *Simulation & Gaming,* 49(4):441-470.

[13] Solinska-Nowak, A., Magnuszewski, P., Curl, M., French, A., Keating, A., Mochizuki, J., Liu, W., Mechler, R., Kulakowska, M., and Jarzabek, L. (2018). An overview of serious games for disaster risk management – Prospects and limitations for informing actions to arrest increasing risk. *International Journal of Disaster Risk Reduction,* 31:1013-1029.

[14] Seo, K., Ryu, H., and Kim, J. (2018). Can serious games assess decision-making biases? Comparing gaming performance, questionnaires, and interviews. *European Journal of Psychological Assessment*. Advance online publication. https://doi.org/10.1027/1015-5759/a000485.

[15] Dunbar, N.E., Miller, C.H., Adame, B.J., Elizondo, J., Wilson, S.N., Schartel, S.G., Lane, B., Kauffman, A.A., Straub, S., Burgoon, J.K., Valicich, J., Bessarabova, E., Jensen, M.L., Jenkins, J., Zhang, J., and Morrison, D. (2014). Mitigating cognitive bias through the use of serious games: Effects of feedback. In: *Persuasive Technology,* Lecture Notes in Computer Science, Spagnolli, A., Chittaro, L., and Gamberini, L. (Eds.), 8462. Cham, Switzerland: Springer.

[16] von Ahn, L., and Dabbish, L. (2008). Designing games with a purpose. *Communication of the ACM,* 51(8):58-67.

[17] P. Smets, P., and Kennes, R. (1994). The transferable belief model. *Artificial Intelligence,* 66:191-234.

[18] Jousselme, A.-L., Boury-Brisset, A.-C., Debaque, B., and Prevost, D. (2014). Characterization of hard and soft sources of information: A practical illustration. In: *Proceedings of the International Conference of Information Fusion*, Salamanca, Spain.

[19] Wang, R.Y., and Strong, D.M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems,* 12 (4):5-34.

[20] Smets, P. (1997). Imperfect information: Imprecision and uncertainty. In: *Uncertainty Management in Information Systems: From Needs to Solutions*, Motro, A., and Smets, P. (Eds.), 225-254. Boston, MA.: Kluwer Academic Publishers.

[21] Klir, G.J., and Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Upper Saddle River, NJ: Prentice Hall International.

[22] International Organization for Standardization. (1994). *ISO 5725-1, Accuracy (Trueness and Precision) of Measurement Methods and Results* – Part 1: General principles and definitions. Technical Report, ISO International Standardization. Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:ed-1:v1:en.



Part III: INTELLIGENCE AND RISK ASSESSMENT UNDER UNCERTAINTY



Chapter 10 – SYSTEMATIC MONITORING OF FORECASTING SKILL IN STRATEGIC INTELLIGENCE

David R. Mandel
Defence Research and Development Canada
CANADA

Better earlier warning than later mourning.

Jewish proverb

As the Jewish proverb conveys, accurate indications about consequential future events that arrive early enough can help decision makers avert trouble. Unsurprisingly, then, forecasting (or prediction, which I use synonymously) plays a vital role in intelligence assessment. According to Allied intelligence doctrine, "analysis does more than look at the current situation, it should be predictive and therefore should address what might happen next, based upon alternative assumptions regarding the actions and reactions of different actors (including the impact of any intervention)" (Ref. [1], §3.38). An effective forecasting capability supports planning and decision making at all levels, ranging from tactical to strategic. And, although the empirical research reported in this chapter focuses on efforts to monitor forecasting accuracy at the strategic level of intelligence production, the issues dealt with apply as well to forecasting at the tactical and operational levels. Moreover, the methods described for monitoring forecast accuracy and forecasters' skill could be applied at those levels as well.

At the strategic level, geopolitical forecasting offers early warning indicators, or what in more recent times has been called "anticipatory intelligence" [2], to key decision makers, such as state leaders and other senior policymakers. Timely, relevant, and accurate geopolitical forecasts promote Allied or national interests by reducing the probability of strategic surprises, mitigating Allied or national security risks, and allowing policymakers to capitalize on opportunities in a dynamic, globalized world. As Sherman Kent noted long ago [3], [4], intelligence that provides the decision maker with timely and accurate indications about the future are among the most important of intelligence products, but they are also the ones that often represent thankless tasks. This, as Kent explained, is because forecasting involves judgement under uncertainty, and analysts don't have exclusive purview over judgements. Policymakers often believe – regardless of whether or not it is true –that they can judge better than analysts. When the intelligence community's judgements concur with their own, they may see in the concordance the goodness of their own judgement. When the judgements are discordant, they may be less than thankful for the dissenting views, especially if they have already committed to a particular course of action or formed a strong bias towards a particular policy option.

10.1 WHY SYSTEMATIC MONITORING MATTERS

Although the intelligence community assigns great importance to the anticipatory or estimative functions of intelligence, the quality of geopolitical forecasting remains largely unverified because intelligence organizations do not routinely and systematically monitor forecasting skill [5], [6], [7], [8]. The decision not to systematically monitor forecasting skill on an ongoing basis with well-established quantitative scoring rules can have many deleterious consequences. Without credible monitoring processes, intelligence organizations simply cannot know how good their forecasts are, how they might be improved, and where improvement is most needed. In the absence of such knowledge and subsequent corrective actions that might have been implemented, intelligence organizations increase the risk of failure to detect strategic threats and opportunities. Without proper monitoring, the intelligence community also remains unnecessarily susceptible to reactive pressure for institutional change – or even institutional erasure in some cases (e.g., Ref. [9]) – after politicized intelligence failures [10], [11]. Without proper monitoring, the intelligence community also

forfeits opportunities for improving forecasting through calibration feedback [12] or recalibration techniques aimed at mitigating biases in judgement, such as overconfidence or underconfidence, that present themselves at the organization level [13]. Nor is the intelligence community well poised to evaluate the effects of training or structured techniques to improve analytic qualities such as forecasting accuracy if such qualities are not carefully measured [14], [15], [16], [17].

Intelligence organizations that wish to track their forecasting accuracy will want to ensure that, at minimum, efforts to support this objective are ecologically valid and methodologically sound. That is, they will want to ensure that the monitoring process represents the analytic environment in which real intelligence forecasts are made, and that, within that context, reasonable efforts are made to promote valid and reliable scoring of forecasting skill characteristics, such as accuracy, discrimination, and calibration. Thus, monitoring systems should ideally track the quality of real forecasting products that are made by real intelligence analysts who are working under normal conditions using well-established quantitative methods that have been applied in other domains of expert judgement under uncertainty (e.g., Refs. [18], [19], [20], [21], [22]).

Such organizations may be less interested in the external validity of the monitoring process insofar as they will be more concerned about their own organization than about how their performance generalizes to other organizations. Nevertheless, concern about the external validity of monitoring procedures should be of interest to the broader intelligence community, and especially to organizations or organizational divisions that are mandated to oversee, coordinate, or provide scientific advice to the wider intelligence community. It is important to know, for instance, whether the organizations that comprise a national or Allied intelligence analytic capability exhibit small or large variance in forecasting skill. Such information sheds light on the important question of whether knowing the performance of one intelligence organization tells us much about the forecasting capability of other intelligence organizations. As we shall see in the research summarized in this chapter, concerns about external validity can be amplified even within a single organization if its units rely on different analytic methods for making forecasts.

10.2 GEOPOLITICAL FORECASTING SKILL

On the strategic level, intelligence forecasting is mainly about geopolitical events. Little is known about expert geopolitical forecasting skill, especially within intelligence organizations. Tetlock's [23] long-term study of close to 300 political experts found evidence of mediocre geopolitical forecasting skill. Experts were highly overconfident, and even the best expert forecasters – namely, the uncertainty-tolerant "foxes," referring to Isaiah Berlin's [24] intellectually playful hedgehog-fox distinction – performed substantially worse than the best statistical models Tetlock tested. Moreover, more seasoned experts fared about the same as those still wet behind the ears. And even more remarkably, experts who forecasted on topics in their areas of expertise did no better than dilettantes – namely, experts who forecasted on topics outside their areas of expertise.

The latter finding has striking implications for intelligence organizations, which routinely organize their expertise along geopolitical fault lines. The presumption of such a system is that an analyst of China will make better forecasts about China than an analyst of Iran, and vice versa. Tetlock's [23] study challenges that bedrock assumption, although it certainly does not invalidate the assumption because the study scores low on ecological validity for two important reasons. First, the experts did not comprise a set of strategic intelligence analysts and, second, the forecasts were made as part of a research study, not as finished intelligence. Experts were neither incentivized or under comparable degrees of accountability pressure. Whether such differences matter remains unclear, but Tetlock's [23] findings should undoubtedly prompt intelligence organizations to redouble efforts to answer such questions.

Many geopolitical topics in Tetlock's [23] study required long-range forecasts that would only be resolved in several years. In contrast, most intelligence forecasts (excluding futures or foresight exercises) are short to medium range and resolve in less than 1 year. A geopolitical forecasting tournament sponsored by the

US government's Intelligence Advanced Research Program Activity (IARPA) elicited forecasts that generally resolved in 1 year or less, thus increasing timeframe comparability to the strategic intelligence realm. The winners of IARPA's Aggregative Contingent Estimation (ACE) program found that elite "superforecasters" could be cultivated using a combination of effective sampling, elicitation, training, and aggregation methods [17], [25], [26]. Superforecasters don't only forecast more accurately, they exhibit superior cognitive abilities [25]. Superforecasters also better discriminate the meaning of verbal probability terms, are less susceptible to content effects on their interpretation of such terms, and are more coherent on other judgement tasks [27]. The findings of the ACE program and other IARPA programs on forecasting suggest several ways in which intelligence organizations could attempt to improve intelligence forecasting, but they do not bring us much closer to knowing how well intelligence organizations currently forecast because, once again, the forecasters are not sampled from intelligence analysts and the forecasts are elicited as part of research rather than on the job.

Ecologically valid studies of geopolitical forecasting skill in intelligence organizations are exceedingly rare. In one example, Lehner *et al.* [28] examined 187 geopolitical forecasts taken from unclassified or declassified intelligence reports and found poor discrimination but fairly good calibration. However, the methods the authors used cast significant doubt on the interpretability of the study's findings. For instance, descriptive statements (e.g., "Arab groups in Kirkuk continue to resist violently what they see as Kurdish encroachment" (see Ref. [28], 730)) were rewritten as forecasts (i.e., "Arab groups in Kirkuk will resist violently what they see as Kurdish encroachment in the January 2007 to July 2009 time frame" (see Ref. [28], 731)). Another critical limitation is that verbal probability qualifiers in the original assessments were omitted in redrafted forecasts. Due to these and other methodological shortcomings, it is difficult to draw conclusions from this research about the quality of strategic intelligence forecasting.

10.3 SYSTEMATIC MONITORING OF FORECASTING SKILL IN CANADA

10.3.1 Initial Phase

The most comprehensive attempt to systematically monitor the forecasting skill in strategic intelligence was conducted in Canada over the last decade. In the initial phase of this effort, Mandel and Barnes [13] conducted a long-term study of geopolitical forecasting skill in strategic intelligence, which examined forecasts extracted from a comprehensive review of 6 years of classified reports produced by the Middle East and Africa (MEA) division of the Intelligence Assessment Secretariat (IAS) in the Canadian government. As part of regular analytic practice in the division investigated, analysts recorded whether their assessments were forecasts or explanatory judgements [29].

Analysts also assigned numeric probabilities to forecasts (but not to explanatory judgements). The nine permissible numeric probabilities – expressed out of 10, these were 0 1, 2.5, 4, 5, 6, 7.5, 9, and 10 – were mapped to verbal probability terms following the lexical standard described in Ref. [29]. For instance, for the probability level 2.5/10 analysts could use the terms *low probability*, *probably not*, or *unlikely*, whereas for 7.5/10 analysts could use *probably*, *probable*, or *likely*. Note that only the verbal probabilities appeared in finished intelligence reports. For example, one forecast (edited to remove sensitive information) was "It is very unlikely [1/10] that either of these countries will make a strategic decision to launch an offensive war in the coming six months." Numeric probabilities appearing in brackets, such as the "1/10" in the example, would not have been printed in final reports. These probabilities were recorded only for auditing and research purposes. In other words, they were not printed in finished intelligence products.

Mandel and Barnes [13] found that roughly three-quarters (74.7%) of intelligence assessments were forecasts. Moreover, just over two-thirds (67.5%) were expressed with sufficient clarity to be quantitatively scored for accuracy. The remaining one-third was excluded from primary analyses either because the description of probability was too vague (i.e., using words such as *might* or *could*) or because the forecasted outcomes were expressed in ways that made them too difficult to code. Among the 1,514 forecasts that were

scored, forecasting skill based on the numeric probabilities that analysts assigned was very good. The mean Brier score, *B*, is a proper scoring rule equal to the mean squared deviation between probabilities assigned to forecasts and outcomes coded 0 for non-occurrence and 1 for occurrence [30], [31]. The Brier score ranges from 0 – 1, with 0 equalling a perfect score. The value of *B* reported in Ref. [13] was 0.074. If forecasts that were excluded due to vagueness were included, *B* expectedly increased, but the mean value remained under 0.10 in various tests of estimate sensitivity.

To put the forecasting skill observed by Mandel and Barnes [13] in more accessible terms, the accuracy rate was 94%. In other words, in 94% of the forecasts, the binary forecast was on the correct side of fifty-fifty, thus pointing decision makers who might rely on such forecasts in the right direction. Using another metric, the normalized discrimination index [32], which computes discrimination over uncertainty (a measure of the proportion of outcome variance explained by the forecasts), Mandel and Barnes [13] found that 76% of outcome variance was explained by the forecasts. Senior analysts also showed better discrimination skill than junior analysts, contrary to what might be inferred from Ref. [23]. Finally, Mandel and Barnes [13] found that, on average, forecasts were underconfident rather than overconfident, once again contrary to what Tetlock [23] observed, but in line with forecasting poll results from the ACE forecasting tournament noted earlier [33].

Underconfidence was significantly greater for forecasts that subject matter experts identified as harder rather than easier, and that the expert coders identified as more rather than less important to intelligence consumers. Such findings are striking in their inconsistency with the well-documented *hard-easy effect* in which harder forecasts tend to produce greater overconfidence than easier forecasts [34]. They suggest that analysts may be expressing uncertainty strategically as a means of pre-emptively deflecting accountability pressure. Indeed, Tetlock and Kim [35] found that when participants were told in advance that they would be accountable for their judgements, they expressed less confidence in their judgements than participants who were either told after their judgements that they would be accountable or who were not told they would be accountable. Couching one's estimates in a high degree of uncertainty is in some sense a form of insurance against reputational injuries. If one happens to be right, there is a minor coverage cost in that one could have been bolder. However, if one turns out to be wrong, then he or she is shielded from blame by the blatant unfalsifiability of a highly uncertain estimate. Analysts – and the intelligence organizations they represent – certainly have more reputational credit to lose from overconfidence than from underconfidence. Perhaps where forecast accuracy is accompanied by high degrees of epistemic uncertainty, analysts may even have more to lose from being well calibrated than underconfident. Although the issue deserves research attention, the preliminary fact pattern is consistent with a social functionalist model of how intelligence analysts are likely to behave when acting strategically as intuitive politicians [36], [37].

Mandel and Barnes [13] further showed that the degree of miscalibration – and underconfidence, more specifically – could be substantially reduced through a simple recalibration procedure in which each probability issued was made one unit more extreme. The values of 0, 0.5, and 1 did not change, but all others became one unit more extreme. For instance, 0.25 became 0.1, just as 0.75 became 0.9. This simple procedure was as effective as more complicated mathematical procedures [38], [39] for improving calibration by reducing underconfidence. Figure 10-1 shows the original model-based calibration curve (with white dots) and the recalibration curve (with black dots) obtained from the extremizing operation just described. Red and blue zones show the domains of overconfidence and underconfidence, respectively. Light grey regions in the blue zone show the benefit of recalibration (underconfidence reduction) and dark grey regions in the red zone show the cost (overconfidence increase).

10.3.2 Second Phase

Monitoring exercises such as the one just described have a peculiar feature in that the forecasting skill assessment is based on numeric probability estimates that intelligence consumers never got to see. Instead, as noted earlier, consumers were presented with vague probability terms, which they were free to interpret

as they please. The question therefore arises, how good are the forecasts when seen through the eyes of a typical consumer? To address this issue, Mandel [8] re-examined the skill of the forecasts reported in Ref. [13] from the perspective of a statistically average intelligence consumer.

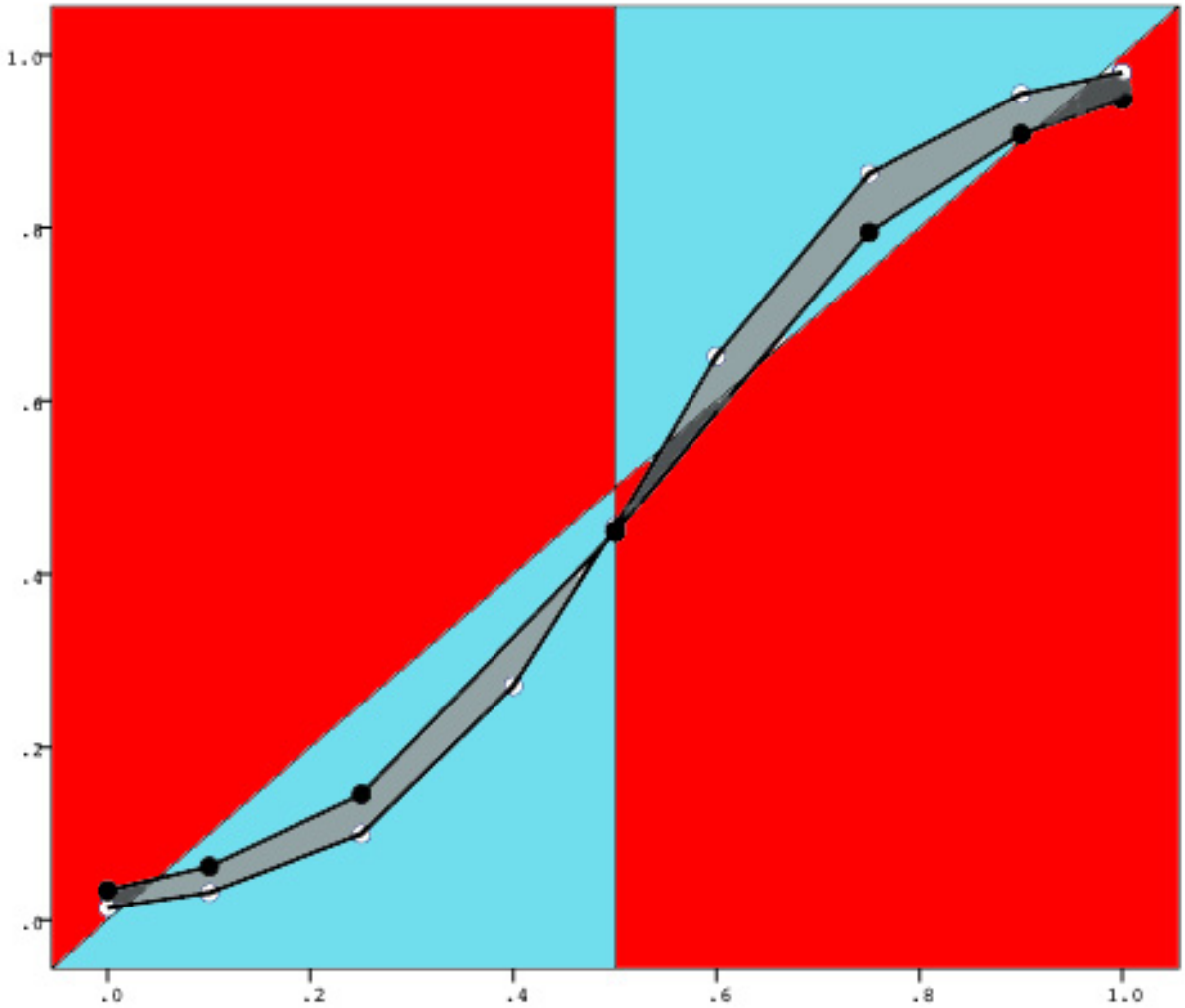


Figure 10-1: Calibration Curve (White Dots) and Recalibration Curve (Black Dots). Adapted from Ref. [13].

Note: The abscissa represents the analyst's forecast probability and the ordinate represents the mean predicted value of a model predicting outcomes from forecasts. Slope = 1 shows the line of perfect calibration.

In the first stage of the research, university students and intelligence analysts were asked to provide their best numeric probability equivalent for each of the 20 probability terms in the lexical standard that had been used by the MEA division [29].

The two subsamples did not statistically differ and their results were combined. Table 10-1 shows the probability terms, their stipulated meaning in terms of numeric probability, and the median numeric probability equivalents obtained in the research. As can be seen in the table, the agreement between the stipulated meanings and average interpretations was very strong. The mean absolute difference between stipulated and median values was equal to 0.019 – an average deviation of less than 2%.

Table 10-1: Probability Terms in the Lexicon and Numeric Probability Equivalents.

Probability Term in Lexicon	Numeric Probability Equivalent		95% CI
	Stipulated	Median	
Will not	0	0	0, 0
No prospect	0	0.02	0, 0.05
Little prospect	0.10	0.14	0.11, 0.15
Extremely unlikely	0.10	0.07	0.05, 0.10
Highly unlikely	0.10	0.10	0.08, 0.10
Very unlikely	0.10	0.10	0.10, 0.12
Low probability	0.25	0.25	0.20, 0.25
Probably not	0.25	0.25	0.20, 0.27
Unlikely	0.25	0.20	0.20, 0.25
Slightly less than even chance	0.40	0.45	0.45, 0.45
Even chance	0.50	0.50	0.50, 0.50
Slightly greater than even chance	0.60	0.55	0.55, 0.55
Probably	0.75	0.75	0.70, 0.75
Probable	0.75	0.71	0.66, 0.75
Likely	0.75	0.75	0.75, 0.80
Highly likely	0.90	0.85	0.85, 0.90
Extremely likely	0.90	0.90	0.90, 0.90
Almost certain	0.90	0.95	0.90, 0.95
Certain	1	1	0.99, 1
Will	1	1	1, 1

Note: Adapted from Mandel [8]. CI = confidence interval. 95% CI based on 1,000 bias-corrected and accelerated bootstrap samples.

In the second stage of the research, participants' median estimates were substituted for analysts' numeric probabilities and the mean Brier score was recomputed. Using the median inferred probabilities, forecasting skill (from the average consumer's perspective) was slightly better (B = 0.071) than that reported in Ref. [13]. This closeness of the two estimates was inevitable given that the standard for mapping verbal to numeric probabilities used in the MEA division was very well matched to median estimates of participants. In other words, forecasting skill was roughly the same from the producers' and average consumer's perspective because the stipulated meanings for terms in the lexicon were very close to the averaged interpretation of consumers. This, in turn, is unsurprising given the care that Barnes took in reviewing the many behavioural studies of verbal probability interpretation that were available at the time he developed the standard. This type of diligent evidence-based analysis is rare in the development of tradecraft standards and practices [6], [14], [40], and the results of Ref. [8] indicate that such effort is worth the investment. It should be commended and emulated in future practice.

10.3.3 Third Phase

The research of Mandel and Barnes [13], [41] and Mandel [8] provides an unparalleled and detailed quantitative analysis of forecasting skill in strategic intelligence. The findings are informative, but also raise many questions. Why was forecasting skill so good? How generalizable are the results? One possibility is that the results were as positive as they were because, as Barnes [29] described, several rather unusual analytic procedures were instituted in the MEA division during the period of study to enhance analytic rigour. Perhaps those practices paid off in terms of better forecasting skill. If so, one might expect that attempts to generalize the findings to other strategic intelligence divisions that did not rely on such procedures but which produced similar products would flounder. Alternatively, as Arkes and Kajdasz [42] posited, general environmental features of intelligence organizations, such as accountability to skeptical stakeholders or requirements to be explicit about judgements and supporting reasoning, might account for the degree of skill observed in the MEA division. In that case, one might expect strategic intelligence forecasting skill to show similar characteristics outside of the MEA division. Yet another hypothesis proposed by Tetlock and Mellers [43] is that the forecasts made by the MEA division were good because they were easy. Although Tetlock and Mellers [43] did not provide empirical evidence in support of their facile-forecasting hypothesis, the hypothesis cannot simply be discounted.

To address these questions and the alternative hypotheses they raise, Mandel and Barnes [44] examined a larger set of strategic intelligence forecasts. The majority of the sample, as in the earlier research, was drawn from the MEA division. However, strategic forecasts from other divisions of the same organization and from team products produced by multiple organizations were also assessed. These forecasts from outside of the MEA division were not generated with the same methods outlined in Ref. [29]. For instance, analysts were not required to categorize their assessments into forecast and non-forecast categories, and they were not required to assign numeric probabilities to forecasts or to use a fixed set of probability terms.

In total, Mandel and Barnes [44] extracted 3,622 forecasts from available intelligence reports, and 73% (2,629) had events that could be unambiguously coded as either having occurred or having not occurred. Of the latter, 77% (2,013) had verbal probability terms for which a numeric probability could be confidently assigned. That sample constituted the forecast dataset. Of the 2,013 forecasts, 1,735 (86%) were in the MEA subsample and 278 (14%) were in the non-MEA subsample. As well, 1,759 (87%) forecasts were from IAS intelligence memoranda and 254 (13%) forecasts were from interdepartmental committees. The vast majority (95%) of the probability terms used in forecasts were terms that appeared in the lexical standard Barnes had devised. This is not surprising given that 87% of the sample was from the MEA division. For these terms, the median numeric probability equivalents from Ref. [8] were substituted. In the remaining 5% of cases, the values were estimated based on similarity to terms in the standard (see Ref. [44] for details).

Addressing the generalizability question first, recall that the mean Brier score was 0.074 in Ref. [13] based on analysts' numeric probabilities, and it was 0.071 in Ref. [8] using inferred numeric probabilities from median sample estimates of the best numeric equivalents to the verbal probability terms. Mandel and Barnes [44] found no significant difference in Brier scores between MEA division and the non-MEA forecasts. Indeed, the two subsamples were virtually indistinguishable: $B = 0.059$ for the MEA subsample and $B = 0.058$ for the non-MEA subsample.

Examining the specific skill components of discrimination and calibration also showed no difference between the groups. Discrimination in both groups was very good, as can be seen in Figure 10-2, which shows the Receiver-Operator Characteristic (ROC) curves for the two subsamples. The ROC curve plots the true-positive hit rate (sensitivity) as a function of the false positive (1 – specificity) rate [45], [46]. The slight difference in area under the curve (0.96 for MEA forecasts and 0.93 for non-MEA forecasts) was not statistically significant.

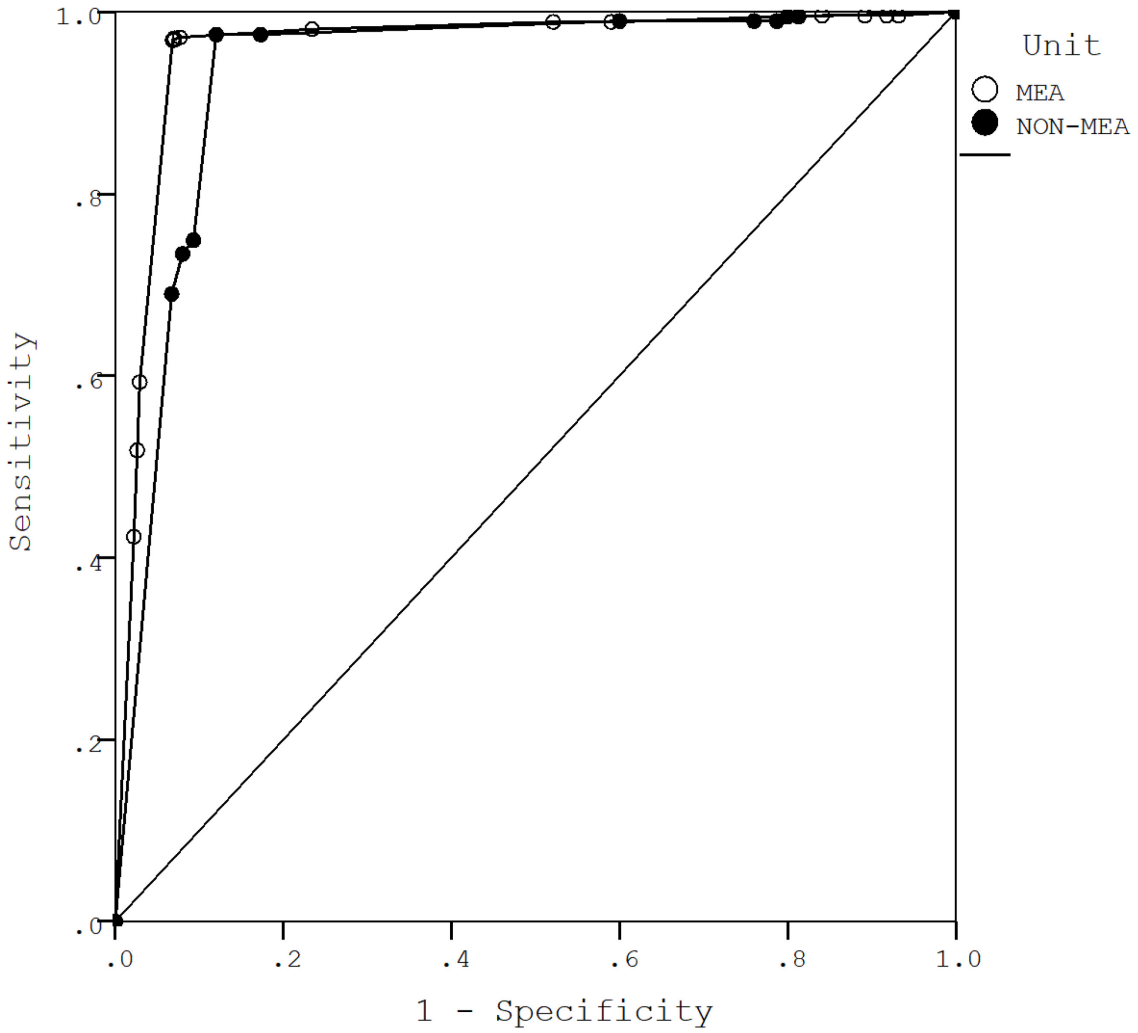


Figure 10-2: Receiver-Operator Characteristic Curves by Unit (from Ref. [44]).

Although the two subsamples did not differ in calibration, and both showed evidence of underconfidence, the degree of underconfidence in the MEA subsample was greater than in the non-MEA subsample. This is evident in Figure 10-3, which shows the model-based calibration curves for the two subsamples. The sigmoidal shape of the curve is more pronounced for the MEA subsample, indicating greater underconfidence. A specific test of underconfidence [34] can be calculated as follows:

$$C_{Conf} = \frac{1}{N_1 + N_2}\left(\sum_{i=1}^{N_1}(o - s) \text{ iff } s < .5 + \sum_{i=1}^{N_2}(s - o) \text{ iff } s > .5\right). \quad (10\text{-}1)$$

In Equation 10-1, o is the outcome (0 or 1) and s is the forecasted probability. Positive values of the measure indicate overconfidence, and negative values indicate underconfidence. In the MEA subsample, the effect of underconfidence was small to medium in size, C_{Conf} = -0.08, $p < 0.001$, Cohen's d = 0.37, whereas in the non-MEA subsample, the effect was very small and marginally significant, C_{Conf} = -0.03, p = .052, Cohen's d = 0.13. MEA forecasts were significantly more underconfident than non-MEA forecasts, but the effect size was small (Cohen's d = 0.23).

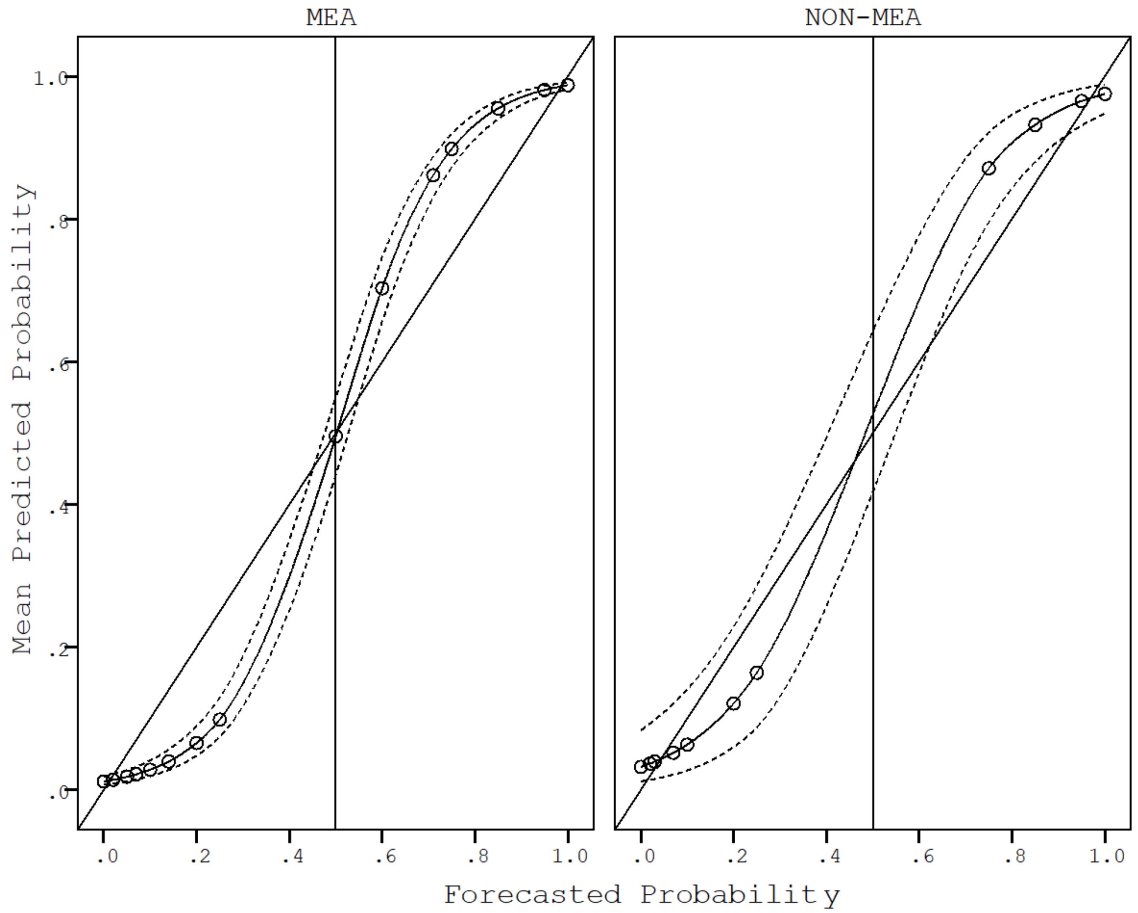


Figure 10-3: Model-Based Calibration Curves by Unit. Dotted lines = 95% CI (from Ref. [44]).

Overall, then, the findings indicate that the forecasting skill characteristics of the MEA division reported in Refs. [13] and [8] are generalizable to other organizational units producing strategic intelligence forecasts. The only notable difference – and again it was small – was the extent of underconfidence, which was greater in the MEA subsample. It is possible that the analytic standards introduced in the MEA division prompted analysts to be more cautious in their assessments, yet this is conjectural. It should be tested in controlled experiments on the effect of accountability processes on expert judgement.

To test the facile-forecasting hypothesis of Tetlock and Mellers [43] – namely, that forecasts were good because they were easy – Mandel and Barnes [44] first compared forecasts from key judgements to forecasts not in key judgements. Key judgements are the most authoritative and informative assessments made in intelligence reports. Therefore, they are less likely than other forecasts to address easy topics that would be uninformative to decision makers. Contrary to the facile-forecasting hypothesis, neither calibration nor discrimination significantly differed between key and non-key forecasts.

In a second test of the hypothesis, Mandel and Barnes [44] compared the 40% of cases in which forecasts were assigned very high certainty (i.e., probabilities greater than or equal to .95 or less than or equal to .05) with the remaining 60% of cases that were less certain (i.e., probabilities greater than .05 but less than .95). The idea here – and one that Tetlock, Mellers and their colleagues have also recently expressed [47] – is that easy forecasts tend to be made with more certainty and, hence, probability values closer to 0 or 1. Contrary to the facile-forecasting hypothesis, neither calibration (in absolute value) nor discrimination significantly differed between the high-certainty and lower-certainty subsamples. Therefore, the facile-forecasting hypothesis was not supported by the findings of Mandel and Barnes [44]. If the facile-forecasting hypothesis

were correct, it would call into question why governments spend vast sums of money tackling prediction problems that are easily or already known. Although the indirect tests of Mandel and Barnes [44] did not support the hypothesis that good forecasting is due to easy topic selection, future research should examine intelligence analysts' forecasting skill under conditions where topic difficulty is either controlled or directly measurable, such as in prediction markets or forecasting tournaments like those sponsored by IARPA.

10.4 CONCLUSION AND RECOMMENDATIONS

The findings of Mandel [8] and Mandel and Barnes [44] demonstrate how intelligence organizations could proactively and quantitatively monitor multiple aspects of their forecasting skill even if they do not use numeric probabilities in forecasts. This demonstration is important because intelligence organizations seldom require forecasts to be made with numeric probabilities. If numeric probabilities were required for monitoring forecasting skill and product accuracy, such a process would currently be infeasible. Moreover, monitoring prediction accuracy would likely remain infeasible for some time to come, given the high degree of institutional resistance to the prospect of communicating probabilistic estimates using numeric probabilities [7], [14], [29], [48], [49]. To illustrate the low-cost feasibility of monitoring forecasting skill, verbal probabilities in the forecast dataset that corresponded to terms in the lexicon described in Ref. [29] were replaced with the median numeric probabilities elicited in Ref. [8]. Verbal probabilities that were not in the lexicon were assigned numeric equivalents on the basis of synonymy to (or, in a few cases, interpolation between) terms in Ref. [29].

The general approach taken by Mandel and Barnes [44] could easily be adopted with other lexicons used by other intelligence organizations or, more generally, by any organization that chooses to communicate uncertainties with verbal probabilities. Of course, the approach just described estimates the forecast accuracy from the statistically average consumer's perspective. It does not estimate the average forecast accuracy based on individual consumers' interpretations of the verbal terms used to qualify probability in forecasts, although that would be easy enough to estimate this from the type of data collected by Mandel [8]. It is quite likely that the average of those mean Brier scores would be substantially worse than the mean Brier score based on the median interpretations of the verbal terms.

As well, the preceding point about the usability of monitoring methods with forecasting systems that use verbal probabilities to characterize uncertainty should not be taken as an endorsement of such an approach to uncertainty. To the contrary, the continued reliance on verbal probabilities and coarse confidence statements is as unacceptable as the arguments in favour of such an approach are flimsy. Most arguments in favour of vague verbiage (a wonderful expression for which Tetlock deserves credit) are not only unsupported by the evidence, but are outright contradicted by it. One argument often heard from intelligence professionals is that their clients would not like to receive estimates with numeric probabilities. This argument relies on no more than hearsay evidence – and probably less, given that such claims are almost invariably devoid of any specifics, which suggest they are largely, if not exclusively, based on proponents' beliefs rather than on what intelligence consumers actually say they prefer. The claim, moreover, is contradicted by scientific studies that have investigated sender and receiver preferences for modes of communicating uncertainty. Those studies indicate that communication senders prefer vague linguistic phrases, whereas communication receivers prefer crisp numeric estimates [50], [51].

Another reason for resisting clarity in the communication of uncertain estimates is that the use of numeric probabilities would convey a false sense of precision or "scientific accuracy." This view reveals a deep misunderstanding of probability (not to mention science). The "false precision" argument is wrong for at least two reasons. First, there is nothing at all scientific about assigning numeric probabilities to events or hypotheses. For instance, if I roll a fair die, I am entitled to say that there is a $1/6^{th}$ probability of it landing on 2, and this has absolutely nothing to do with science, despite its numeric expression. Likewise, I can forecast, as I have recently done in an experimental prediction market, that Benjamin Netanyahu will have a 20% chance of

being indicted by March 31, 2018. Of course, in early March of 2018, I would have been much more confident in the first estimate than the second (at the time of writing this March 31 has passed and Netanyahu was not indicted), and that brings me to the second reason for rejecting the false precision argument. Simply put, numeric probabilities can be expressed precisely or imprecisely. Had there been an option for expressing uncertainty in my forecast regarding Netanyahu, I would have placed a fairly large confidence interval around the estimate. The point is that it is still far better to say that there is a 60 – 80 % chance that *x* will occur (or to say that there is a 70% chance plus or minus 10%, or something similar) than to say it's *likely* that *x* will occur, as the former numeric expressions, although imprecise, are neither vague nor ambiguous. If analysts would develop the habit of assessing a numeric probability range, they would not need to have a vague verbal probability lexicon supplemented by another vague and coarse confidence indication (typically no more than a "low, moderate, or high" rating, as in NATO intelligence doctrine [1]).

A more probable reason for the entrenchment of verbal probabilities as a means of communicating uncertainty is that intelligence professionals as intuitive politicians believe that their best interests are served by having the interpretive wiggle room that vague verbiage affords [52]. Yet, there too, belief appears to be contradicted by scientific evidence. Consider a recent experiment by Jenkins *et al.* [53]. Participants were first given probabilistic predictions in numeric (e.g., "20% likelihood"), numeric-verbal (e.g., "20% likelihood [unlikely]"), verbal-numeric (e.g., "unlikely [20% likelihood]"), or verbal (e.g., "unlikely") formats. Then, they were informed that the prediction was erroneous. Participants viewed the numeric estimates as least incorrect and the verbal estimates as most incorrect. Assessments of the forecaster's credibility showed a similar pattern. If these findings are generalizable, intelligence organizations might inadvertently be making themselves easier targets of blame-game tactics when so-called intelligence failures become media firestorms and political hot potatoes [10], [11].

Even if intelligence consumers preferred numeric probabilities to verbal uncertainty expressions and the use of the former did not put analysts qua intuitive politicians at a disadvantage, defenders of the status quo would still be entitled to ask what evidence there is that using numeric probabilities would yield any real benefit to assessment accuracy. The first line of evidence can be drawn directly from the research already reviewed in this chapter. Mandel and Barnes [13], [44] found that a substantial proportion of strategic intelligence forecasts were uncodable because they were expressed with the vague terms that did not lend themselves to credible probability ranges – for instance, words such as *might* or *could*. Using numeric probability estimates with or without ranges would overcome this problem, and the probabilities assessed would be transparent to intelligence consumers – and verifiable – in all cases. Second, Friedman *et al.* [47] have recently shown that if the granularity of forecasts made with numeric probabilities on a 101-point scale is reduced by making the forecasts match the midpoint values of stipulated ranges in lexicons (such as those used by the US and the UK [54], [55]), forecasting accuracy is also substantially reduced. Moreover, the reduction in accuracy was greatest for the best forecasters, which indicates that the current system of using verbal terms to communicate uncertainties is especially harmful to the quality of the analytic cream of the crop.

To sum up, although the methods for monitoring the quality of intelligence forecasts in intelligence organizations that were developed by Mandel and Barnes [13], [44] could be applied to forecasts produced under current production systems, ideally, those systems should be changed if the intelligence community wants to optimize its assessment accuracy and communication fidelity. If the intelligence community is serious about its professional commitment to cultivating predictive accuracy in intelligence assessments for its decision-making customers, it must cast aside its arcane practices that lack evidential support. At minimum, it should adhere to the following two recommendations:

1) Intelligence organizations should continually monitor their forecasting performance, taking steps to ensure that monitoring practices are rigorous and not susceptible to cognitive and motivational biases. Monitoring processes should support intelligence accountability and organizational learning.

2) Intelligence organizations should revamp their approaches to communicating probability and confidence as facets of uncertainty in judgements. Specifically, they should quantify estimates of uncertainty, even if those estimates are based on judgements requiring subjectivity.

Quantification of uncertainty and subjectivity in intelligence assessment are not inconsistent, and it is time intelligence organizations realize and accept this fact.

Here we can anticipate one more predictable rejoinder from defenders of the status quo – the point that, while prediction may be a part, even an important part, of intelligence assessment, it is not all there is. Intelligence involves many other forms of assessment and must also be narratively compelling in order to engage the decision-making reader. This specific form of straw-man argument might as well be called *argumentum ad poeticum* in honour of Sherman Kent's [3] famous poet-mathematician distinction. The claim is no more than a straw-man argument because absolutely no one denies that intelligence assessments are comprised of more than only predictions. This is obvious. However, defenders of vague verbiage will surely be uncomfortable with the finding reported earlier that fully three-quarters of judgements made in strategic intelligence reports were predictive in nature [13]. No amount of poetic flair will compensate for clarity and verifiability in this hefty proportion of assessment cases. Decision makers deserve no less, even if they have been perennially lackadaisical in pressing for what they need.

10.5 REFERENCES

[1] North Atlantic Treaty Organization (2016). *Allied Joint Doctrine for Intelligence Procedures AJP-2.1*. Brussels, Belgium: NATO Standardization Office.

[2] Clapper, J. (2014). *The National Intelligence Strategy of the United States of America: 2014*. Washington DC: Office of the Director of National Intelligence.

[3] Kent, S. (1964). Words of estimative probability. *Studies in Intelligence*, 8(4):49-65.

[4] Kent, S. (1994). Estimates and influence. In: *Sherman Kent and the Board of National Estimates: Collected Essays,* Steury, D.P. (Ed.), 51-59. Washington DC: Center for the Study of Intelligence.

[5] Betts, R.K. (2007). *Enemies of Intelligence: Knowledge and Power in American National Security*. New York, NY: Columbia University Press.

[6] Dhami, M.K., Mandel, D.R., Mellers, B.A., and Tetlock, P.E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science,* 10(6):753-757.

[7] Friedman, J.A., and Zeckhauser, R. (2016). Why assessing estimative accuracy is feasible and desirable. *Intelligence and National Security,* 31(2):178-200.

[8] Mandel, D.R. (2015). Accuracy of intelligence forecasts from the intelligence consumer's perspective. *Policy Insights from the Behavioral and Brain Sciences,* 2(1):111-120.

[9] Moynihan, D.P. (1991, May 19). Do we still need the C.I.A.? The State Dept. can do the job. *The New York Times*, E17.

[10] Johnson, L.K. (2007). A shock theory of congressional accountability for intelligence. In: *Handbook of Intelligence Studies,* Johnson, L.K. (Ed.), 343-360. New York, NY: Routledge.

[11] Tetlock, P.E., and Mellers, B.A. (2011). Intelligent management of intelligence agencies: Beyond accountability ping-pong. *American Psychologist*, 66(6):542-554.

[12] Rieber, S. (2004). Intelligence analysis and judgmental calibration. *International Journal of Intelligence and CounterIntelligence,* 17(1):97-112.

[13] Mandel, D.R., and Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences of the United States of America,* 111(30):10984-10989.

[14] Chang, W., Berdini, E., Mandel, D.R., and Tetlock, P.E. (2018). Restructuring structured analytic techniques in intelligence. *Intelligence and National Security,* 33 (3):337-356.

[15] Chang, W., and Tetlock, P.E. (2016). Rethinking the training of intelligence analysts. *Intelligence and National Security,* 31(6):903-920.

[16] Mandel, D.R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Frontiers in Psychology*, 6:387.

[17] Mellers, B.A., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S.E., Moore, D., Atanasov, P., Swift, S.A., Murray, T., Stone, E., and Tetlock, P.E. (2014). Psychological strategies for winning geopolitical forecasting tournaments. *Psychological Science,* 25(5):1106-1115.

[18] Åstebro, T., and Koehler, D.J. (2007). Calibration accuracy of a judgmental process that predicts the commercial success of new product ideas. *Journal of Behavioral Decision Making,* 20(4):381-403.

[19] Goodman-Delahunty, J., Granhag, P.A., Hartwig, M., and Loftus, E.F. (2010). Insightful or wishful: Lawyers' ability to predict case outcomes. *Psychology, Public Policy and Law* 16 (2):133-157.

[20] Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes,* 39(1):98-114.

[21] Murphy, A.H., and Winkler, R.L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association* 79 (387):489-500.

[22] Lin, S.-W., and Bier, V.M. (2008). A study of expert overconfidence. *Reliability Engineering and System Safety,* 93(5):711-721.

[23] Tetlock, P.E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.

[24] Berlin, I. (1953). *The Hedgehog and the Fox: An Essay on Tolstoy's View of History.* New York, NY: Weidenfeld & Nicolson.

[25] Mellers, B.A., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., and Tetlock, P.E. (2015). Identifying and cultivating "Superforecasters" as a method of improving probabilistic predictions. *Perspectives in Psychological Science,* 10(3):267-281.

[26] Tetlock, P.E., and Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. New York, NY: Crown Publishing Group.

[27] Mellers, B.A., Baker, J.D., Chen, E., Mandel, D.R., and Tetlock, P.E. (2017). How generalizable is good judgment? A multi-task, multi-benchmark study. *Judgment and Decision Making,* 12(4):369-381.

[28] Lehner, P., Michelson, A., Adelman, L., and Goodman, A. (2012). Using inferred probabilities to measure the accuracy of imprecise forecasts. *Judgment and Decision Making,* 7(6):728-740.

[29] Barnes, A. (2016). Making intelligence analysis more intelligent: Using numeric probabilities. *Intelligence and National Security,* 31(3):327-344.

[30] Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review,* 78:1-3.

[31] Murphy, A.H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology* 12:595-600.

[32] Yaniv, I., Yates, J.F., and Smith, J.E.K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin,* 110(3):611-617.

[33] Atanasov, P., Rescober, P., Stone, E., Swift, S.A., Servan-Schreiber, E., Tetlock, P., Ungar, L., and Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, 63(3):691-706.

[34] Lichtenstein, S. and Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance,* 20(2):159-183.

[35] Tetlock, P.E., and Kim, J.I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology,* 52(4):700-709.

[36] Tetlock, P.E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review*, 109(3):451-471.

[37] Mandel, D.R., and Tetlock, P.E. (2016). Debunking the myth of value-neutral virginity: Toward truth in scientific advertising. *Frontiers in Psychology* 7:451.

[38] Turner, B.M., Steyvers, M., Merkle, E.C., Budescu, D.V., and Wallsten, T.S. (2014). Forecast aggregation via recalibration. *Machine Learning* 95 (3):261-289.

[39] Baron, J., Mellers, B.A., Tetlock, P.E., Stone, E., and Ungar, L.H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis* 11 (2):133-145.

[40] Mandel, D.R. (2019). Can decision science improve intelligence analysis? In: *Researching National Security Intelligence: A Reader*. Coulthart, S., Landon-Murray, M., and Van Puyvelde, D. (Eds.). Multidisciplinary Approaches, 117-140. Washington, DC: Georgetown University Press.

[41] Mandel, D.R., Barnes, A., and Richards, K. (2014). *A Quantitative Assessment of the Quality of Strategic Intelligence Forecasts*. DRDC Toronto Technical Report 2013-036. Toronto, ON: DRDC.

[42] Arkes, H.R., and Kajdasz, J. (2011). Intuitive theories of behavior. In: *Intelligence Analysis: Behavioral and Social Scientific Foundations,* Fischhoff, B., and Chauvin, C. (Eds.), 143-168. Washington DC: The National Academies Press.

[43] Tetlock, P. and Mellers, B. (2014). Judging political judgment. *Proceedings of the National Academy of Sciences of the United States of America,* 111(32):11574-11575.

[44] Mandel, D.R., and Barnes, A. (2018). Geopolitical forecasting skill in strategic intelligence. *Journal of Behavioral Decision Making* 31 (1):127-137.

[45] Swets, J.A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin,* 99(2):181-198.

[46] McClelland, G.H. (2011). Use of signal detection theory as a tool for enhancing performance and evaluating tradecraft in intelligence analysis. In: *Intelligence Analysis: Behavioral and Social Scientific Foundations*, Fischhoff, B., and Chauvin, C. (Eds.), 83-99. Washington DC: The National Academies Press.

[47] Friedman, J.A., Baker, J.D., Mellers, B.A., Tetlock, P.E., and Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly*, 62(2):410-422.

[48] Marchio, J. (2014). "If the weatherman can…": The intelligence community's struggle to express analytic uncertainty in the 1970s. *Studies in Intelligence* 58 (4):31-42.

[49] Spielmann, K. (2016). I got algorithm: Can there be a Nate Silver in intelligence? *International Journal of Intelligence and CounterIntelligence,* 29(3):525-544.

[50] Brun, W., and Teigen, K.H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes,* 41(3):390-404.

[51] Wallsten, T.S., Budescu, D.V., Zwick, R., and Kemp, S.M. (1993). Preferences and reasons for communicating probabilistic information in numeric or verbal terms. *Bulletin of the Psychonomic Society*, 31(2):135-138.

[52] Tetlock, P.E., Mellers, B.A., and Scoblic, J.P. (2017). Bringing probability judgments into policy debates via forecasting tournaments. *Science,* 355(6324):481-483.

[53] Jenkins, S.C., Harris, A.J.L., and Lark, R.M. (2019). When unlikely outcomes occur: The role of communication format in maintaining credibility. *Journal of Risk Research*, 22(5):537-554.

[54] Ho, E., Budescu, D.V., Dhami, M.K., and Mandel, D.R. (2015). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science & Policy* 1(2):43-55.

[55] Irwin, D., and Mandel, D.R. (2018). *Methods for Communicating Estimative Probability in Intelligence to Decision-Makers: An Annotated Collection*. DRDC Scientific Letter DRDC-RDDC-2018-L017. Toronto, ON: DRDC.



Chapter 11 – INFORMATION SECURITY CONTINUOUS MONITORING (ISCM) – PRIORITIZING RISKS IN DEFENSIVE CYBER OPERATIONS (DCO)

Akhilomen O. Oniha and Greg Weaver
Army Research Laboratory
UNITED STATES

The ability for commanders to know and understand an organizational attack surface, its vulnerabilities, and associated risks is a fundamental aspect of command decision making. In the cyberspace domain, ongoing monitoring sufficient to ensure and assure effectiveness of security controls related to systems, networks, threat, and cyberspace, by assessing security control implementation and organizational security status in accordance with organizational risk tolerance and within a reporting structure designed to make real-time, data-driven risk management decisions is paramount.

11.1 INTRODUCTION TO ISCM FOR DCO

The National Institute of Standards and Technology (NIST) Special Publication (SP) 800-137 [1], Information Security Continuous Monitoring (ISCM) for Federal Information Systems and Organizations, defines continuous monitoring as "maintaining ongoing awareness to support organizational risk decisions." NIST SP 800-137 further defines information security continuous monitoring as "maintaining ongoing awareness of information security, vulnerabilities, and threats to support organizational risk management decisions."

While the definitions mentioned above were not formalized until September 2011, the concept of monitoring has (and continues) to be recognized as sound management practice throughout the United States (US) Federal Government. For instance:

- In 1997, Office of Management and Budget (OMB) Circular A-130 [2], Appendix III, required agencies to review and monitor their information systems' security controls and to ensure that system changes do not have a significant impact on security, that security plans remain effective, and that security controls continue to perform as intended.
- In 2002, the Federal Information Security Management Act (FISMA) of 2002 further emphasized the importance of continuously monitoring information system security by requiring agencies to conduct assessments of security controls at a frequency appropriate to risk, but no less than annually.
- In 2011, OMB issued memorandum M-11-33 [3], FY 2011 Reporting Instructions for the Federal Information Security Management Act and Agency Privacy Management. The memorandum provided instructions for annual FISMA reporting and emphasized monitoring the security state of information systems on an ongoing basis with a frequency sufficient to make ongoing, risk-based decisions.
- In 2013, OMB memorandum M-14-03 [4], FY 2014, Page 1, Enhancing the Security of Federal Information and Information Systems, further specified, the requirement to manage information security risk on a continuous basis includes the requirement to monitor the security controls in Federal information systems and the environments in which those systems operate on an ongoing basis – one of six steps in the NIST Risk Management Framework. This allows agencies to maintain ongoing awareness of information security, vulnerabilities, and threats to support organizational risk management decisions.

The aforementioned requirements are supported by the following additional NIST guidance:

- **NIST SP 800-37 Rev. 2 [5]**, Guide for Applying the Risk Management Framework (RMF) to Federal Information Systems: A Security Life Cycle Approach provides guidelines for applying the RMF to federal information systems to include conducting the activities of security categorization, security control selection and implementation, security control assessment, information system authorization, and security control monitoring. This guideline further describes monitoring security controls at the system level (RMF Step 6) and also includes an organization-wide perspective, integration with the System Development Life Cycle (SDLC), and support for ongoing authorizations.
- **NIST SP 800-39 [6]**, Managing Information Security Risk: Organization, Mission, and Information System View, provides guidance for an integrated, organization-wide program for managing information security risk to organizational operations (i.e., mission, functions, image, and reputation), organizational assets, individuals, other organizations, and the Nation resulting from the operation and use of federal information systems. This guideline provides a structured, yet flexible approach for managing information security risk that is intentionally broad-based, with the specific details of assessing, responding to, and monitoring risk on an ongoing basis provided by other supporting NIST security standards and guidelines. The guidance provided in this publication is not intended to replace or subsume other risk-related activities, programs, processes, or approaches that organizations have implemented or intend to implement addressing areas of risk management covered by other legislation, directives, policies, programmatic initiatives, or mission/business requirements. Rather, the information security risk management guidance described herein is complementary to and can be used as part of a more comprehensive Enterprise Risk Management (ERM) program.
- **NIST 800-137** further expands the concepts presented in NIST SP 800-39, which describes three key organization-wide ISCM activities (i.e., monitoring for effectiveness, monitoring for changes to systems and environments of operation, and monitoring for compliance) and NIST SP 800-37 Rev. 1, which describes monitoring security controls at the system level (RMF Step 6) and also includes an organization-wide perspective (integration with the System Development Life Cycle [SDLC] and support for ongoing authorizations), in order to provide guidelines sufficient for developing an ISCM strategy and implementing an ISCM program.

11.1.1 The Risk Management Framework (RMF) vs. Security Authorization vs. ISCM

In essence, the RMF developed by NIST, describes a disciplined and structured 6-step process that integrates information security and risk management activities into the system development life cycle. Security authorization and ISCM are integral components of the RMF; security authorization occurs in RMF Step 5, the Authorize step, and ISCM occurs in RMF Step 6, the Monitor step. NIST developed the RMF to provide a more flexible, dynamic, approach for effective management of information system-related security risk in highly diverse environments and throughout the system development life cycle. Both security authorizations and ongoing monitoring are critical components of the RMF. The six RMF steps are shown in Table 11-1.

As previously mentioned, security authorization occurs in RMF Step 5, the Authorize step. Security authorization is the process by which a senior management official, the Authorizing Official (AO), reviews security-related information describing the current security posture of an information system and uses that information to determine whether or not the mission/business risk of operating a system is acceptable – and if it is, explicitly accepts the risk.

Table 11-1: RMF Steps.

RMF Approach	
Step	**Description**
Step 1. *Categorize*	Categorize the information system and the information processed, stored, and transmitted by that system based on an impact analysis.
Step 2. *Select*	Select an initial set of baseline security controls for the information system based on the security categorization, tailoring and supplementing the security control baseline as needed based on an organizational assessment of risk and local conditions.
Step 3. *Implement*	Implement the security controls and describe how the controls are employed within the information system and its environment of operation.
Step 4. *Assess*	Assess the security controls using appropriate assessment procedures to determine the extent to which the controls are implemented correctly, operating as intended, and meeting the security requirements as described in the system security plan.
Step 5. *Authorize*	Authorize information system operation based on a determination of the risk resulting from the operation of the information system, and the decision that this risk is acceptable.
Step 6. *Monitor*	Monitor and assess selected security controls in the system on an ongoing basis including assessing security control effectiveness, documenting changes to the system or environment of operation, conducting security impact analyses of the associated changes, and reporting the security state of the system to appropriate organizational officials.

A security authorization can be the initial authorization, ongoing authorization, or a reauthorization as described in Table 11-2.

Correspondingly, ISCM occurs in RMF Step 6, the Monitor step. The intent of ISCM is to give organizational officials access to security-related information on demand, enabling timely risk management decisions, including ongoing security authorization decisions. To accomplish, the ISCM framework needs to support the monitoring of applicable security controls with the frequency needed to provide AOs with the necessary and sufficient information to make effective, risk-based decisions, whether by automated or procedural/manual means.

The guideline also describes an effective ISCM program as one implemented with the appropriate rigor and assessment frequencies to support the organization's mission/business requirements, risk tolerance, and security categorization, is essential to establishing an OA process. Ongoing risk determinations and risk acceptance decisions by senior leaders depend on having relevant and credible near-real-time information to help inform such risk determinations and decisions. As shown in Table 11-3, NIST SP 800-137 defines six steps for achieving effective ISCM.

Table 11-2: Security Authorization Categories.

Security Authorization Category	Description
Initial Authorization	Initial authorization is defined as the initial (start-up) risk determination and risk acceptance decision based on a zero-base review of the information system conducted prior to the Information System's entering the operations/maintenance phase of the system development life cycle. The zero-base review includes an assessment of all security controls (i.e., system-specific, hybrid, and common controls) contained in a security plan and implemented within an information system or the environment in which the system operates.
Ongoing Authorization (OA)	OA is defined as the subsequent (i.e., follow-on) risk determinations and risk acceptance decisions taken at agreed upon and documented frequencies in accordance with the organization's mission/business requirements and organizational risk tolerance. OA is a time-driven or event-driven security authorization process whereby the AO is provided with the necessary and sufficient information regarding the near-real-time security state of the information system (including the effectiveness of the security controls employed within and inherited by the system) to determine whether or not the mission/business risk of continued system operation is acceptable.
Reauthorization	Reauthorization is defined as the static, single-point-in-time risk determination and risk acceptance decision that occurs after initial authorization. In general, reauthorization actions may be time-driven or event-driven; however, under OA, reauthorization is typically an event-driven action initiated by the AO or directed by the Risk Executive (function) in response to an event that drives information security risk above the previously agreed upon organizational risk tolerance. Reauthorization is a separate activity from the ongoing authorization process, though security-related information from the organization's ISCM program may still be leveraged to support reauthorization. Note also that reauthorization actions may necessitate a review of and changes to the ISCM strategy which may in turn, affect ongoing authorization.

Table 11-3: ISCM Steps.

ISCM Approach	
Step	**Description**
Step 1. *Define an ISCM Strategy*	Define an ISCM strategy based on risk tolerance that maintains clear visibility into assets, awareness of vulnerabilities, up-to-date threat information, and mission/business impacts.

ISCM Approach	
Step	**Description**
Step 2. *Establish an ISCM Program*	Establish an ISCM program determining metrics, status monitoring frequencies, control assessment frequencies, and an ISCM technical architecture.
Step 3. *Implement an ISCM Program*	Implement an ISCM program and collect the security-related information required for metrics, assessments, and reporting. Automate collection, analysis, and reporting of data where possible.
Step 4. *Analyze*	Analyze the data collected and Report findings, determining the appropriate response. It may be necessary to collect additional information to clarify or supplement existing monitoring data.
Step 5. *Respond*	Respond to findings with technical, management, and operational mitigating activities or acceptance, transference/sharing, or avoidance/rejection.
Step 6. *Review and Update*	Review and Update the monitoring program, adjusting the ISCM strategy and maturing measurement capabilities to increase visibility into organizational assets and awareness of vulnerabilities, further enable data-driven control of the security of an organization's information infrastructure, and increase organizational resilience.

11.1.2 The Current US Government Approach for ISCM

11.1.2.1 ISCM for Non-High-Impact Systems: DHS CDM Program

In the US, to support the implementation of OMB M-14-03, the Department of Homeland Security (DHS) made available the Continuous Diagnostics and Mitigation (CDM) Program. Since FY 2013, the DHS CDM has provided agencies with *non-high-impact systems* the opportunity to use a suite of tools and capabilities to identify cybersecurity risks on an ongoing basis, prioritize these risks based on potential impacts, and enable cybersecurity personnel to mitigate the most significant problems first. These tools include sensors that perform automated searches for known cyber vulnerabilities, the results of which feed into dashboards that alert network managers, enabling agencies to allocate resources based on the risk. DHS also provides a federal dashboard-related infrastructure. The tools and services delivered through the CDM program are to provide the ability to enhance and automate existing agency continuous network monitoring capabilities, correlate and analyze critical security-related information, and enhance risk-based decision making at agency and federal levels.

11.1.2.2 ISCM for High-Impact Systems: Agency-Specific (e.g., DOD) ISCM Program

Agencies with high-impact systems are not required to participate in the DHS CDM Program (e.g., the Department of Defense (DOD), is one of the 18 agencies with high-impact systems). Largely, agencies with *high-impact systems* have already established advanced ISCM capabilities and solutions. Most, have leveraged products/tools provided through General Services Administration's acquisition vehicle and/or researched and developed their own Government off the Shelf solutions to complete installation of agency dashboards, and monitored attributes of authorized users operating in their agency's computing environment.

11.2 ISCM RISK SCORING METHODOLOGY (NOTIONAL)

11.2.1 Compliance-Based vs. Performance-Based vs. Risk-Based

The following section describes the evolution of Army Research Laboratory (ARL) ISCM risk scoring methodology. As shown in Figure 11-1, to achieve its objective to deliver persistent situational awareness across multiple tiers, mission areas of operations, and security domains and support informed and actionable risk management decisions, ARL has gradually and methodically evolved its risk scoring strategy from a compliance-based to a risk-based scoring.

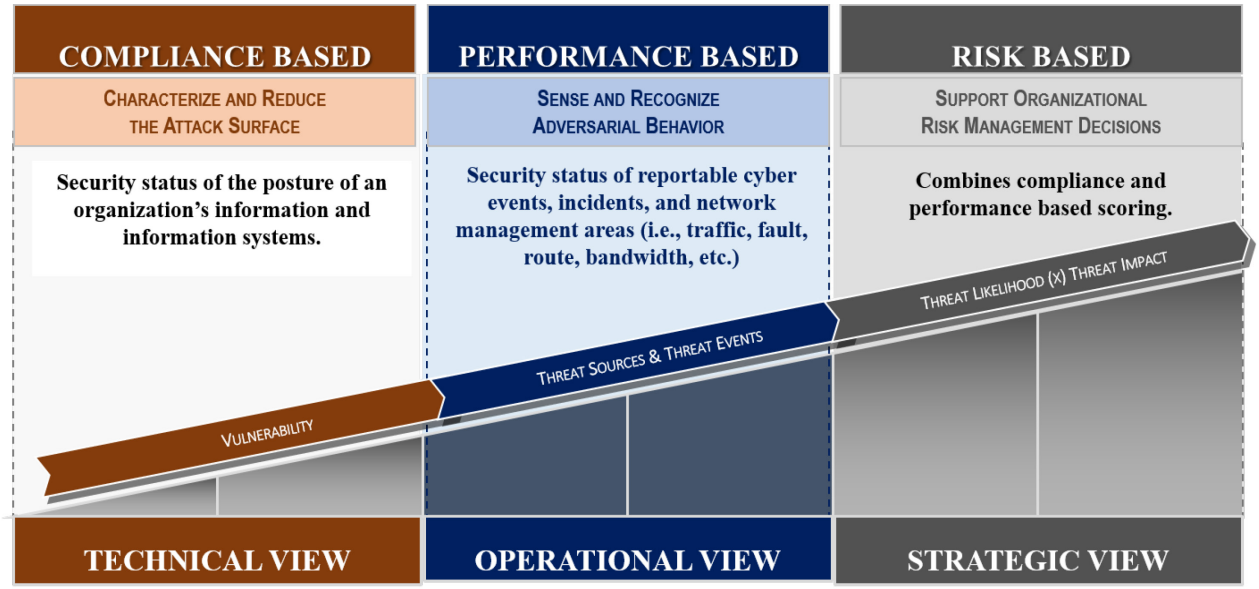


Figure 11-1: ARL ISCM Risk Scoring Strategy.

11.2.2 Dynamic-State Risk Posture

ARL's risk scoring approach provides for a dynamic-state risk posture that incorporates a series of views, such that each view fits into a scheme that provides a comprehensive "snapshot in time" that explains the environment from both a threat/vulnerability perspective and a risk/impact perspective. These views are described as follows:

- **Technical View:** ARL's ISCM solution support to risk assessment is *compliance-based* and leverages the identification of technical vulnerabilities both internal and external to organizations.
- **Operational View:** As the ARL's ISCM solution continues to mature, its support to risk assessment will become *performance-based* adding (to the *compliance-based* baseline), the identification of information network operations management areas to include traffic, fault, performance, bandwidth, route, and other network management areas as well as the identification of relevant cybersecurity incidents and events.
- **Strategic View:** Fully developed, ARL's ISCM solution support to risk assessment will become *risk-based* adding (to the *performance-based* baseline), threats to organizations; impact (i.e., harm) to organizations that may occur given the potential for threats exploiting vulnerabilities; and likelihood that harm will occur.

Overall, the ARL's risk scoring construct provides an enterprise operational view captured in user-defined dashboards, complete with decision engines, which aids operators and decision makers understand and make decisions regarding anomalies associated with their organizational areas of responsibility.

The two initial functions of the Risk Management Widget (RMW) are risk identification and risk scoring. The functionality that determines the cause of the risk, identifies issues that are mitigated in order to eliminate a particular risk at the asset or site level. Issues are prioritized by their influence on risk. The relative risk scoring functionality uses vulnerability discovery results to estimate risk where risk score is based on the exposure of vulnerable services to external networks. Risk scores are presented by site, asset, or vulnerability (Figure 11-2).

The current ARL ISCM solution which is built atop of a big data platform and initially includes five individual widget/analytic capabilities, working in conjunction with one another to provide a dynamic cyber hunting capability and enhanced decision support via a risk categorization and prioritization capability. As illustrated in Figure 11-3, those individual capabilities include asset management, antivirus compliance, network management, vulnerability management and risk management. The first four capabilities serve as the building blocks to generate the risk picture in the fifth capability.

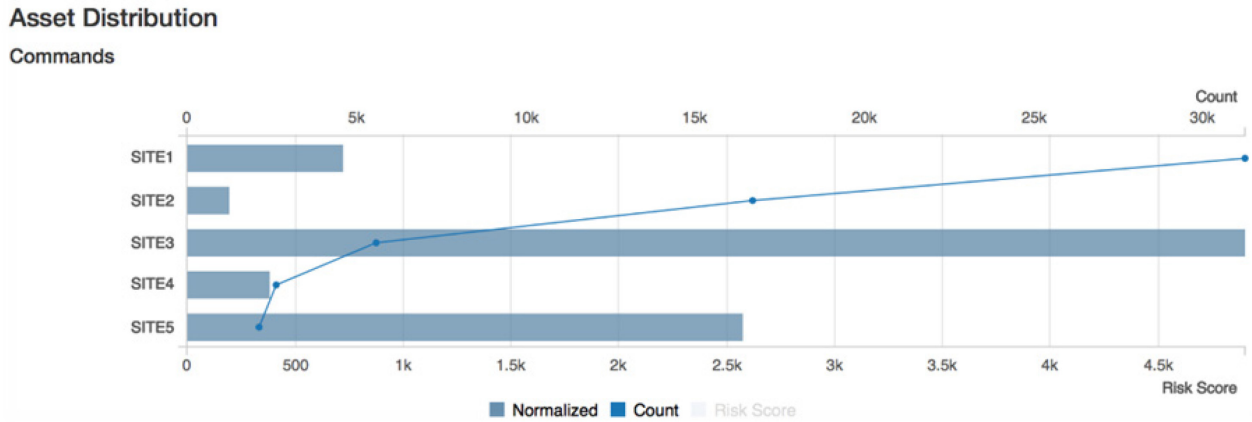


Figure 11-2: ISCM Risk Scoring User Interface.



Figure 11-3: ISCM Risk Scoring Widgets.

ISCM's current approach to risk management satisfies its initial goal of aiding stakeholders in the comparison and prioritization of higher-risk vs. lower-risk assets. However, when looking at assets independently, the risk score does not provide the context necessary to assess the actual risk of an asset being compromised. The ARL team is actively investigating and pursuing the implementation of a probabilistic risk scoring widget for future integration.

The ARL team began by leveraging the generic algorithm for cyber risk, where risk is a function of threats, vulnerabilities and impact, $R = f(T\ x\ V\ x\ I)$. We supplemented the equation with additional characteristics that were critical to the defensive operations mission, including:

- **Confidence Level:** The belief that an asset is exposed to a particular vulnerability by taking into account all relevant observations (i.e., output from all tools: HBSS, ACAS, etc.). This value is a derived percentage, currently implemented using term frequency-inverse document frequency algorithm [7] and represents the certainty that the host in question actually has the factors deemed to be vulnerable (i.e., software version, patch version, operating system, etc.).
- **Threat Multiplier:** A factor associated with the exposure of a vulnerability to an external network for remote exploitation. Vulnerabilities exposed to a wide area network and remotely exploitable, produce the highest threat multiplier.
- **Temporal Certainty Multiplier:** A factor associated with the age and freshness of the vulnerability scan reports. As scan information ages, the potential risk from vulnerabilities that cannot be confirmed as mitigated increases. The temporal certainty multiplier represents the increase as a factor, which is multiplied against the vulnerability instance risk score. The time period is determined by comparing the greatest last seen of all risk factors to current time.
- **Exposure Duration Multiplier:** A factor indicating how long an unresolved risk was first detected. The time period is determined by comparing the earliest first seen of all risk factors to current time.
- **Exploit Threat Multiplier:** A dynamic factor that allows a security analyst to amplify/decrease the weighting of a CVE risk score throughout the system. The value is set on the user interface through a RESTful service, which updates the entity model. The current factor values can be set to be very low; low; moderate; high; and very high; depending upon the level of exploitation or other threat intelligence.

11.3 TOP TEN CHALLENGES: ISCM AND RISK SCORING METHODOLOGY

11.3.1 Challenge #1 – Managing the Magnitude of the Attack Surface to be Monitored

The nature and origin of this challenge derives primarily from the magnitude of the attack surface to be continuously monitored.

For example, the DOD is an immense and complex organization; active users of DOD Information System (IS) and Platform Information Technology (PIT) systems and networks include more than 1.4 million men and women serving on active duty, 750 thousand civilian personnel, and 1.1 million National Guard and Reserve members. In addition, the DOD's presence extends globally to more than 145 countries, 6,000 locations, and 600,000 building and structures. Overall, the DOD mission requires an Information Technology (IT) and Cybersecurity workforce of over 170,000 persons to support day-to-day operations to more than 15,000 classified and unclassified networks and 7 million computers and IT devices worldwide.

Overtime, the magnitude of the DOD's attack surface has made it difficult to monitor. Equally challenging to manage, the enormous size of the DOD has made it an easier target for a growing number of cyber-attacks, which are becoming more frequent, sophisticated, aggressive, and dynamic. Every day, the largest attack surface in the world, represented by IS and PIT systems and networks, are subject to serious cyber-attacks and exploitation activities – any of which could, without adequate implementation of cybersecurity controls, have adverse effects on mission operations, assets, individuals, our nation, and our allies.

11.3.2 Challenge #2 – Prioritizing the Asset and Data (Elements and Sets) to Be Monitored

As shown in Figure 11-4, given the complexity of information systems and networks, a discussion of *what* and *how* to monitor is challenging unless a common framework is employed.

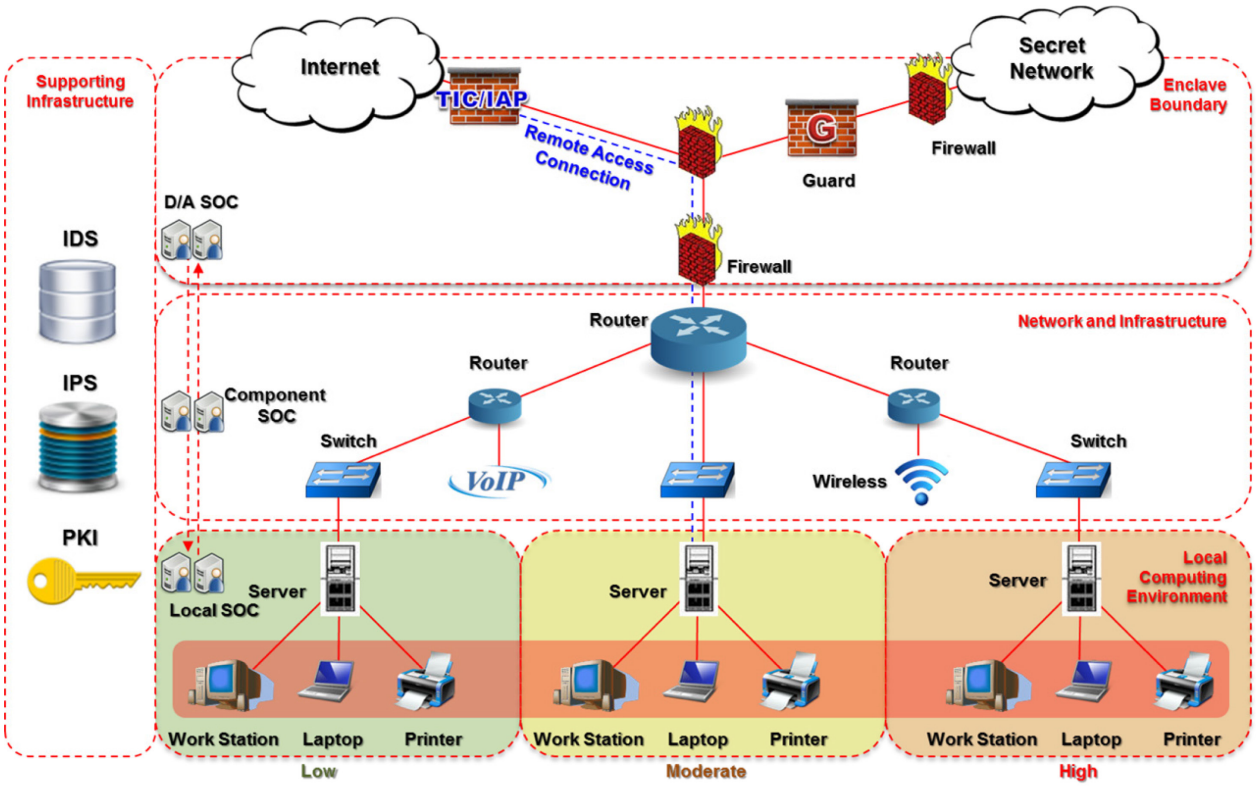


Figure 11-4: What to Continuously Monitor.

In the US, for example, most DOD organizations leverage the IATF to partition the IA technology aspects of information systems into four areas:

1) Local Computing Environment;
2) Enclave Boundary;
3) Network and Infrastructure; and
4) Supporting Infrastructure.

By partitioning the discussion into these four areas, the monitoring aspects of IA technology for an information system can be prioritized, focused upon, and more clearly presented.

The Local Computing Environment typically contains servers, clients, printers, and the applications installed on them. Applications include, but are not limited to, those that provide services such as scheduling or time management, printing, word processing, or directories.

The Network and Infrastructure provides connectivity between enclaves and contains the information transmission components to move information between the network nodes. Network and infrastructure components include routers, switches, wireless, Voice over Internet Protocol (VoIP), etc.

The Enclave Boundary is the point at which information enters or leaves the organizational enclave. Most organizations have extensive connections to networks outside their control. Therefore, monitoring this layer of protection is needed to ensure that the information entering does not affect the organization's operation or resources, and that the information leaving is authorized. This layer of protection is usually established through firewalls, guards, remote access connections, etc.

Supporting Infrastructure provides the foundation upon which IA mechanisms are used in the network, enclave, and local computing environments for securely managing the system and providing security-enabled services. Supporting Infrastructure provides security services for networks, end-user workstations, web applications, and file servers, and single-use infrastructure machines (e.g., higher-level Domain Name Server [DNS] services, and higher-level directory servers). Applicable components to be monitored include, for example, Intrusion Detection/Prevention Systems (IDS/IPS) and Public Key Infrastructure (PKI).

Once assets have been prioritized, the challenge is determining the ISCM data (elements and sets) to be collected to ensure:

- Data collection requirements are standardized.
- Data sets and data elements to be collected are aligned to support applicable mission areas of operations.
- The right data is collected once and reused many times to support informed and actionable risk management decisions in multiple mission domains.
- Synergism with data analysis and data presentation requirements.

11.3.3 Challenge #3 – Defining Monitoring Frequencies

To support ongoing authorization, security-related information for all implemented security controls, including inherited common controls, needs to be monitored, generated, and collected at the frequency specified in the organizational ISCM strategy. Security-related information may be collected using automated tools or via other methods of assessment depending on the type and purpose of the security control and the desired degree of rigor.

The challenge here is two-fold:

- First, automated tools may not generate security-related information sufficient to support the ongoing authorization in making risk determinations because:
  - Additional assurance is needed;
  - The tools do not generate information for every implemented security control or every part of an implemented control; or
  - The tools do not generate information on specific technologies or platforms.
- Second, in such cases, manual or procedural security control assessments need to be conducted at the organizationally defined frequencies to cover any gaps in automated security-related information generation. The procedurally generated assessment results are usually difficult to complete in the manner determined appropriate by the organization.

11.3.4 Challenge #4 – Integrating Existing ISCM Capabilities

In order to drive the synergistic integration and implementation of the ISCM program, organizational challenges include:

- Examining and determining the variance between currently deployed vs. to-be-deployed ISCM capabilities (e.g., tools, mechanisms, programs, etc.);
- Classifying and categorizing current capability gaps and/or domain areas in need of improvement;
- Integrating and eliminating redundancies and duplicative efforts;
- Prioritizing and defining minimum capability requirements in increments;

- Determining milestones and timelines for incremental ISCM implementation; and
- Aiming to lock step with advances in security automation that can provide organizational-wide efficiencies of scale.

11.3.5 Challenge #5 – Prioritizing Security Automation Domains

A security automation domain is an information security area that includes a grouping of tools, technologies, and data. To support ISCM, data within domains needs to be captured, correlated, analyzed, and reported to present the cybersecurity status of the organization that is represented by the domains monitored. Examples of security automation domains include (not all inclusive): Vulnerability Management, Patch Management, Event Management, Incident Management, Malware Detection, Asset Management, Configuration Management, Network Management, License Management, Information Management, and Software Assurance.

The challenge when (and after) prioritizing security automation domains is identifying the tools and technologies to better support ISCM while providing standardized specifications that enable the interoperability and flow of data between these domains.

11.3.6 Challenge #6 – Standardizing Data Analysis and Presentation Requirements

The challenge includes standardizing data analysis and presentation that consider:

- The context in which data is analyzed and presented to support mission, tier, domain, and/or consumer view requirements;
- The lifecycle methods, tools, techniques, and processes to support the analytics process;
- The supporting capabilities for the standard examination and interpretation of integrated data collected (e.g., data reduction, data correlation, data mining, data visualization); and
- The standardization of data analysis and data presentation requirements to drive systemic changes and support informed and actionable risk management decisions.

11.3.7 Challenge #7 – Defining Continuous Monitoring Dashboard Requirements

The challenge includes defining dashboard requirements that consider (not all inclusive):

- Normalization, consolidation, correlation, visualization, and presentation of Continuous Monitoring data;
- Fusion of metrics, measures, and analytics into continuous monitoring dashboards;
- The business logic that allows the organization to prioritize finding and remediation actions based on their risk environment while still meeting minimum standards;
- Scalability requirements; and
- Data strategy to securely interconnect, aggregate, integrate, and interoperate with other dashboards.

11.3.8 Challenge #8 – Defining ISCM Reporting Requirements

Reporting requirements do not drive the ISCM strategy but play a role in the frequency of monitoring. For instance, if OMB policy requires quarterly reports on the number of unauthorized components detected and corrective actions taken, the organization would monitor the system for unauthorized components at least

quarterly. The challenge is to provide a reporting solution with the ability to tailor output and drill down from high-level, aggregate metrics to system-level metrics; and allow for data consolidation into Security Information and Event Management (SIEM) tools and dashboard products.

11.3.9 Challenge #9 – Establishing Performance Benchmarks

The challenge includes determining meaningful, understandable, and measurable ISCM metrics and measures that adequately:

- Explains the criticality, complexity, and uniqueness of the ISCM problem space (across multiple tiers and mission domains);
- Measures accurately, consistently, and in a reproducible fashion within each tier/domain;
- Promotes consistent analysis (collected ISCM data is collated and presented using standard calculations, comparisons, and presentations within each tier/domain); and
- Promotes common reporting (ISCM findings and conclusions are reported clearly, trends are identified and explained, and comparisons made to promote systemic changes within each tier/domain).

The measures and supporting metrics to be used are classified as one of the four categories of measurements:

- Implementation: Measures adherence of continuous monitoring implementation requirements throughout the organization;
- Compliance: Measures adherence to technical and non-technical policies, architectural standards, configuration requirements, or processes mandated for the operation of continuous monitoring;
- Performance: Measures the use of ISCM standard procedures, tactics and techniques, and the overall execution of operational tasks; and
- Impact: Measures operational effects attributable to the network or system on the performance of specific continuous monitoring mission-related tasks and processes, in terms of delays, reductions in accuracy or reliability of information, and excess effort/resources required to offset these effects.

Metrics associated with measuring maturity of a capability or used for risk scoring will be provided as part of the requirements reported to the DOD or Federal level dashboard or as part of the Federal scoring.

Metrics should be designed to facilitate decision making and improve performance and accountability through collection, analysis, and reporting of relevant performance-related data. Continuous monitoring metrics can be realistically obtained and useful for performance improvements depending on the maturity of the security control implementation. Adherence to metrics requirements will ensure the data being collected, measured, and analyzed is:

- Measured accurately, consistently, and in a reproducible fashion (data is collected using standard units and terms in a way that permits multiple measurements of the same or similar events using common methods and standards); and
- Analyzed consistently (the collected data is collated and presented using standard calculations, comparisons, and presentations).

11.3.10 Challenge #10 – Collecting, Reusing, and Sharing of Data

Create a robust and integrated collect-once, reuse-many capability framework that maximizes the reuse of data, information, and knowledge across multiple mission domains (avoiding duplication of efforts and providing accurate and timely information to decision makers).

An organization's ISCM vision can only be realized by a future state where transparent, open, agile, timely, relevant, and trusted data and information sharing occurs to promote freedom of manoeuvrability across the information environment. Organizations must continue to overcome the challenges and improve established methods for data and information sharing across internal and external boundaries; these methods will take into account and remain agile to accommodate differing levels of trust based on the environment, situation, and extended enterprise.

11.4 ENHANCING INTELLIGENCE ASSESSMENT BY MEANS OF AN ISCM FRAMEWORK

The employment of the Big Data Platform (BDP) [8] to ingest, aggregate, correlate and enrich cyber data from a variety of sources and provide an integrated interface or dashboard view, enables commanders and mission owners to make higher-confidence decisions based upon historical, trended, or other data sources. The current design of the ARL ISCM tool is based on a flexible entity model that allows for the inclusion of new data elements from a variety of feeds. One feed type of particular interest that has yet to be integrated into ISCM is threat intelligence. The ISCM team is actively exploring the notion of extending the entity model to include threat intelligence feeds from a variety of sources, because understanding the threats that can exploit active vulnerabilities is clearly a better measure of an organizations risk posture than just vulnerabilities alone.

The caveat for integrating threat intelligence is that there are an innumerable number of threat intelligence sources and no standard or common taxonomy for such feeds. In order to optimally integrate these feeds into ISCM, some mechanism for normalization and/or standardization across feeds needs to occur. Some manual knowledge engineering and feed analysis needs to take place to accomplish this task.

The initial solution may be as simple as reviewing the various threat feeds and extracting the attributes of interest that are common across all the feeds and useful for identifying where "badness" exists.

The inclusion of threat intelligence feeds affords us the ability to automate the adjustment of our threat multiplier and the exploit threat multiplier parameters in ISCM, currently a manual process based upon the analysis of the indicator by a human analyst. Automated intelligence feeds permit the dynamic adjustment to our ISCM risk scoring methodology, which will be highlighted through an updated defensive cyber operations dashboard.

For example, the DOD Cyber community has further expanded the definition of ISCM to support the ongoing observation, assessment, analysis, and diagnosis of an organization's cybersecurity: posture[1], hygiene[2], and operational readiness[3]. This is to reflect the effects of ISCM in support of risk management decisions within different mission areas of operation to include:

- **Risk Management (Framework)[4]:** This mission area includes the ongoing monitoring of the posture of information systems and networks, decreasing the level of effort of our current assessment and authorization process (static and manually intensive every three years) and enables ongoing authorization. By using ISCM, the risk management framework provides the means for prioritizing remediation based upon its relevant organizational risk impact (e.g., technical and operational).

[1] For the purpose of this document, cyber posture means the extent to which the residual risk resulting from cyber hygiene and operational readiness is appropriate given the sensitivity of the business operation and the cost of maintaining that level of hygiene and posture.

[2] For the purpose of this document, cyber hygiene means (a) removing (some or all) known cyber weaknesses from the system, and/or (b) reducing the attack surface of the system.

[3] For the purpose of this document, cyber operational readiness means the strength of an organization's ability to find and mitigate/accept (as appropriate) cyber weaknesses.

[4] A structured approach used to oversee and manage risk for an enterprise.

- **Network Operations:** This mission area includes the ongoing monitoring of network management areas to include traffic, fault, performance, bandwidth, route, and other network management areas (e.g., vulnerability management, account management).
- **Cyber Defence[5]:** This mission area includes the ongoing monitoring of adversarial behavior and its impacts on operations (i.e., monitor and understand intrusions, attack sensing and warning, indications and warning, advanced persistent threats, and other signs of cyber-attack and exploitation activities).

11.4.1 Is It Possible for the Intelligence Community to Benefit from a Methodology Similar to ISCM?

While we continue development of ISCM for cyber defence operations to enrich our operational view through automated feeds, we propose that it is possible for the intelligence community to benefit from a continuous monitoring methodology and "scoring" approach similar to ISCM for initial triage or analysis of source data, but would require extensive mission analysis and capabilities to be developed to identify sources, automate ingests, and integrate data feeds for scoring and presentation in a continuous monitoring approach.

11.4.2 Benefits of Implementing a Continuous Monitoring Framework to Enhance Intelligence Assessments

The proposed construct needed for the benefits of implementing an ISCM framework to support the Intelligence mission area of operation may include (not all inclusive):

- **Promotes Informed and Actionable Risk Management Decisions:** The value of ISCM is not in the promotion of a particular course of action, but rather in its ability to facilitate continuous compliance status and situational awareness of organizational assets. As a capability, an ISCM framework may help the intelligence community prioritize risk mitigation options and strategies by distinguishing the technical and operational impacts within the larger mission context aiding management in decision making and operations.
- **Empowers Leaders and Improves Organizational Accountability:** By providing the knowledge needed to support informed and actionable risk management decisions (affecting their area of responsibility), ISCM helps to drive behavior and to empower leaders, down to the lowest operational levels, to be responsible owners and account for their information systems and networks.
- **Simplifies Regulatory Compliance Through Integrated Data Management:** Through a collect-once and reuse-many framework, the intelligence community may eliminate duplicative data call efforts, accelerate time to mission execution, enable enterprise-wide reporting efficiencies, and simplify regulatory compliance reporting requirements through integrated data management and automation.
- **Fosters Resource Recovery of People, Operations, and/or Technology:** By automating defined ISCM processes and eliminating redundancy, intelligence organizations will free up cyber and intelligence operators. These resources could potentially be reallocated to perform other technical and operational level tasks that cannot be automated and minimize non-technical processes reducing (and in some areas eliminating) the human factor and errors.
- **Delivers persistent Situational Awareness (SA):** ISCM delivers persistent situational awareness across multiple tiers, mission areas of operations, and security domains. As envisioned, ISCM provides for a dynamic-state risk posture that incorporates a series of views such that each view fits

[5] For the purpose of this document, Cyber Defense represents the actions taken to defend an organization's information against unauthorized activity.

into a scheme that provides a comprehensive "snapshot in time" that explains the environment from not only a pure threat/vulnerability perspective, but also a risk/impact one. This is a technical construct that provides an enterprise operational view captured in user-defined dashboards complete with decision engines, which aid the operators and decision makers in understanding and making decisions regarding anomalies associated with their organizational areas of responsibility and/or an enterprise as a whole. This persistent situational awareness provides a level of confidence that enables decision makers to be proactive, rather than reactive, in their decision making.

11.5 REFERENCES

[1] NIST. Special Publication (SP) 800-137, "Information Security Continuous Monitoring (ISCM) for Federal Information Systems and Organizations", https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-137.pdf. (30 September 2011).

[2] OMB Circular A-130, "Managing Information as a Strategic Resource", https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/OMB/circulars/a130/a130revised.pdf. (28 July 2016).

[3] OMB. Memorandum M-11-33, "FY 2011 Reporting Instructions for the Federal Information Security Management Act and Agency Privacy Management", https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2011/m11-33.pdf. (14 September 2011).

[4] OMB. Memorandum M-14-03, "Enhancing the Security of Federal Information and Information Systems", https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-03.pdf. (18 November 2013).

[5] NIST. Special Publication (SP) 800-37 Revision 2, "Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy", https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-37r2.pdf. (20 December 2018).

[6] NIST. Special Publication (SP) 800-39, "Managing Information Security Risk: Organization, Mission, and Information System View", https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-39.pdf. (01 March 2011).

[7] TF-IDF. "Term Frequency-Inverse Document Frequency". (n.d.) Retrieved from http://www.tfidf.com/. (18 July 2016).

[8] Bart, D.V., Big Data Platform (BDP) and Cyber Situational Awareness Analytic Capabilities (CSAAC). Retrieved from http://www.disa.mil/~/media/Files/DISA/News/Conference/2016/AFCEA-Symposium/4-Bart_Big-Data_Platform_Cyber.pdf. (2016, April 22).



Chapter 12 – BOOSTING INTELLIGENCE ANALYSTS' JUDGMENT ACCURACY: WHAT WORKS, WHAT FAILS?[1, 2]

David R. Mandel
Defence Research and Development Canada
CANADA

Christopher W. Karvetski
KaDSci, LLC
UNITED STATES

Mandeep K. Dhami
Middlesex University
UNITED KINGDOM

12.1 INTRODUCTION

Intelligence organizations routinely call upon their analysts to make probability judgments and test hypotheses under conditions of uncertainty. These expert judgments can inform important policy decisions concerning national and international security. Traditionally, analysts have been expected to accumulate domain expertise and apply this along with critical thinking skills to arrive at timely and accurate assessments for decision makers. In the US, developers of analytic tradecraft (i.e., the methods developed within the intelligence community to support its analytic functions) such as Richards Heuer Jr. and Jack Davis introduced so-called "Structured Analytic Techniques" (SATs) to support the analyst in the assessment process, but these methods were largely optional tricks of the trade. That state of affairs changed following two notable geopolitical events (i.e., the September 11, 2001, terrorist attacks by Al Qaeda and the 2003 invasion of Iraq) that were attributed in part to striking intelligence failures. These events prompted reviews of the intelligence community with ensuing organizational reforms that, among other things, aimed at debiasing intelligence analysts' judgments [1]. In the US, the Intelligence Reform and Terrorism Prevention Act of 2004 mandated the use of SATs in intelligence production and SATs became a staple topic in most analytic training programs [2], [3], [4]. Much the same set of organizational reforms was enacted in other Western countries such as the UK [5].

Although the number of SATs has skyrocketed over the last decade [6], [7], as others have lamented in recent years [2], [8], [9], [10], there has been little effort to test their effectiveness. Instead, most SATs have been adopted on the basis of their perceived face validity with the belief that, although imperfect, they must be better than nothing. At the same time, the intelligence community has rarely considered using post-analytic techniques to improve judgment [11]. For instance, Mandel and Barnes [12] showed that intelligence analysts' strategic forecasts were underconfident, but that much of this bias could be eliminated by recalibrating their judgments to make them more extreme [13], [14]. Similarly, the accuracy of probability judgments can be improved post-judgment by recalibration so judgments respect one or more coherence principles, such as the axioms of probability calculus – a statistical process called coherentization [15]. Karvetski *et al.* further observed that weighting individuals' contributions to aggregated judgments also improved accuracy above the gains achieved using an unweighted arithmetic average. In the present research, we examine the accuracy of intelligence analysts' probability judgments in an experimental task. We examine the effectiveness of ACH as well as recalibration and aggregation methods with the aim of addressing the prescriptive question: what works – and what fails – to improve judgment accuracy?

[1] This chapter is reprinted with permission from Ref. [16].

[2] Funding support for this work is provided by Ref. [17].

12.2 THE ANALYSIS OF COMPETING HYPOTHESES TECHNIQUE

The Analysis of Competing Hypotheses [7], [18] is one of the most widely known SATs, and one of only nine that is listed in the US Government's Tradecraft Primer [19], [20]. The US Government describes ACH as a diagnostic technique whose main function is to externalize analytic hypotheses and evidence. It further claims that ACH helps analysts overcome common cognitive biases, such as primacy effects, confirmation bias, and other forms of premature cognitive closure that can undermine the accuracy of forecasts or other probabilistic assessments. The US Government also asserts that ACH "has proved to be a highly effective technique when there is a large amount of data to absorb and evaluate" [19], yet it does not cite any evidence to support that claim. The UK handbook conveys comparable exuberance for ACH, noting, "The approach is designed to help analysts consider all the evidence in the light of all the hypotheses as objectively as possible" (see Ref. [20], 14).

ACH includes several steps, but the core of the tradecraft method involves generating a matrix in which Mutually Exclusive and (preferably) Collectively Exhaustive (MECE) hypotheses are listed in columns and pieces of relevant evidence are listed in rows. The analyst then assesses the consistency of each piece of evidence with each hypothesis starting on the first row and moving across the columns. For each cell, the analyst rates evidence hypothesis consistency on a 5-point ordinal scale (i.e., -2 = highly inconsistent, -1 = inconsistent, 0 = neutral or not applicable, 1 = consistent, 2 = highly consistent). However, only the negative scores (-1 and -2) are tallied for each hypothesis. For instance, if there were five pieces of information (i.e., five rows) and Hypothesis A had ratings {2, 2, 2, 2, -2} and Hypothesis B had ratings {0, 0, 1, 0, -1}, Hypothesis B with an inconsistency score of −1 would be rated as more likely to be true than Hypothesis A with a score of -2. In other words, ACH requires that analysts disregard evidential support for hypotheses in the information integration process. This feature of the method may have been motivated by a misapplication of Popper's [21] ideas about the merits of falsification as a strategy for scientific discovery. Popper's claim that hypotheses could only be falsified but never proven pertained to universal hypotheses such as "all swans are white" because a single non-white swan is sufficient to disprove the claim. Most hypotheses of interest in intelligence, however, are not universal but rather deal with events in a particular context (e.g., Iran is developing a nuclear weapon), and few could be falsified outright by a single disconfirming piece of evidence [22].

ACH also includes a subsequent evidential editing phase: once the matrix is populated with consistency ratings, the analyst is encouraged to remove evidence that does not appear to differentiate between the alternative hypotheses. However, there is virtually no guidance on how such assessments of information usefulness should be conducted. For instance, the US Government merely instructs, "The 'diagnostic value' of the evidence will emerge as analysts determine whether a piece of evidence is found to be consistent with only one hypothesis or could support more than one or indeed all hypotheses. In the latter case, the evidence can be judged as unimportant to determining which hypothesis is more likely correct" (see Ref. [18], 15). The UK handbook is more precise, stating "For each hypothesis ask the following question: 'If this hypothesis were true, how likely would the evidence be?'" (see Ref. [7], 15; [20]). Yet, it vaguely advises analysts to "pay most attention to the most diagnostic evidence – i.e., that which is highly consistent with some hypotheses and inconsistent with others" (see Ref. [20], 17). If evidence is subsequently disregarded, then analysts are expected to recalculate the sum of the negative (inconsistency) ratings. These scores are then meant to reflect the rank ordering of hypotheses by subjective probability, with the hypothesis receiving the smallest inconsistency score being judged as most likely to be true in the set of hypotheses being tested.

ACH is not a normative method for probabilistic belief revision or hypothesis testing, but it has become an institutionalized heuristic that intelligence organizations have deemed to be effective without compelling reasons or evidence [2], [11], [22], [23], [24], [25]. As already noted, ACH disregards useful information about evidential support for hypotheses and it requires analysts to self-assess information utility without providing a clear definition of utility, let alone a computational method for estimating such utility. Perhaps even more fundamental is the omission of a clear definition of consistency, which could signify a range of

meanings, such as the probability of the evidence given the hypothesis, the probability of the hypothesis given the evidence, the plausibility or the necessity of one given the other, or simply a subjective sense of the representativeness of one to the other – namely, the representativeness heuristic [26]. In addition, ACH does nothing to ensure that analysts consider prior probabilities or objective base rates when revising their beliefs about hypotheses in light of new evidence. In sum, there are many reasons to be skeptical about the effectiveness of ACH.

Unfortunately, there is little scientific research on ACH, and what exists must be interpreted cautiously for several reasons, such as small sample sizes [27], [28], [29], lack of control groups [27] or appropriate control groups [29]. Moreover, virtually all published studies have omitted critical, quantitative measures of judgment accuracy, focusing instead on distal considerations such as whether ACH reduces (the highly equivocal notion of) "confirmation bias" [30]. Yet, despite the many serious limitations of research on ACH (and SATs, more generally), the intelligence studies literature has shown little concern regarding the lack of adequate research to support the widespread use of SATs, including ACH. Rather, a recent review article concluded that ACH was "found to be effective and had a highly credible evidence base ..." (see Ref. [3], 377). This conclusion is unwarranted not only because of the methodological weaknesses noted earlier, but also because the extant findings are at best equivocal. For instance, whereas Lehner *et al.* [28] find that ACH reduced confirmation bias in non-analysts, it had no effect on analysts.

12.3 THE PRESENT RESEARCH

A central aim of our research was to examine how the accuracy and logical coherence of intelligence analysts' judgments about the probability of alternative MECE hypotheses depended on whether or not analysts were trained in and used ACH on the experimental task. In addition to this SAT, we also explored the value of statistical post-judgment methods for improving expert judgment, such as recalibrating experts' probabilities in ways that remedy certain coherence violations (i.e., non-unitarily and/or non-additivity), and by aggregating experts' judgments using varying group sizes and weighting methods.

We tested the effectiveness of ACH by randomly assigning intelligence analysts from the same population to experimental conditions that either used ACH or did not. One group of analysts was recently trained to use ACH as part of their organization's training and they were required to use ACH on the experimental task. The other group of analysts was drawn from the same analytic cohort (i.e., same organization and taking the same training course) but they were not instructed to use ACH (or any SAT for that matter) and were not exposed to ACH training until after the experiment was completed. The task, which involved a hypothetical scenario, required analysts to assess the probabilities of four MECE hypotheses that corresponded to four tribes in a region of interest. Participants were asked to assess the probabilities that a detained individual (i.e., the target) from the local population belongs to each of the four tribes. Participants were given the tribe base rates and diagnostic conditional probabilities for 12 evidential cues (e.g., "speaks Zimban"), along with the cue values (6 present and 6 absent) for the target. Furthermore, two tribes (Bango and Dengo, hereafter B and D) were grouped as friendly (F), whereas the other two (Acanda and Conda, hereafter A and C) were grouped as hostile (H).

If ACH proponents' claims about the technique's effectiveness are warranted, we should find greater probabilistic judgment accuracy in the ACH condition than in the control condition. As noted earlier, to the best of our knowledge, there is no clear evidence to support the claim that ACH improves probabilistic judgment accuracy. Indeed, one non-peer-reviewed study that compared various degrees of ACH support (e.g., ACH on its own or with additional training) across experimental groups found that accuracy was best among those participants in the no-ACH control group [31]. However, insufficient information was provided to interpret these results with any confidence.

In addition, if proponents' claims about the effectiveness of ACH in promoting soundness of judgment are true, we might expect to find that analysts recently trained in and aided by ACH produce probability judgments that are more coherent than those unaided by ACH. We tested this proposition by examining the degree to which probability judgments in both groups respect the axioms of unitarity and additivity. To do so, we drew on predictions of support theory, a non-extensional descriptive account of subjective probability which posits that one's probability judgments are a function of his or her assessments of evidential support for a focal hypothesis and its alternative [32], [33]. Support theory predicts an unpacking effect, in which the sum of the probabilities assigned to a MECE partition with more than two subsets of an event, x, exceeds P(x). Unpacking effects have been shown in several studies [32], [33], [34], [35]. For instance, in two experiments with undergraduate participants, Mandel [36] found that the mean unpacking factor – namely, the ratio of the sum of unpacked probability estimates to the packed estimate – was 2.4 comparing forecasts of terrorism (i.e., the packed forecast) to forecasts of terrorism unpacked into acts committed by Al Qaeda or by operatives unaffiliated with Al Qaeda. No research has yet examined whether intelligence analysts' probability judgments are susceptible to the unpacking effect. In the present research, the unpacking effect would be observed if P (A) + P (C) > P (H) and/or P (B) + P (D) > P (F). According to the additivity axiom, these inequalities should be equalities, given that $A \cap C = \emptyset$ and $A \cup C \equiv H$; likewise, $B \cap D = \emptyset$ and $B \cup D \equiv F$.

Extending our investigation into the coherence of analysts' probability judgments, we further tested whether analysts' judgments respect the unitarity axiom, which states that the probabilities assigned to a MECE set of hypotheses should sum to unity. Support theory predicts that partitions of a sample space into more than two subsets will yield an unpacking effect. Thus, in the present research, support theory predicts P (A) + P (B) + P (C) + P (D) > 1.0, in violation of the unitarity axiom, which requires that these probabilities sum to unity. The unitarity axiom also requires that the binary complements P (H) and P (F) sum to 1.0, although support theory predicts agreement with the axiom in the case of binary complements. Some studies find agreement with support theory's prediction for binary complements [32], [37], [38], whereas others find that the sum of the probabilities people assign to binary complements is less than unity [39], [40], [41], [42]. Consistent with the latter studies, Mandel [43] found that intelligence analysts who were given a series of binary classification tasks to complete provided total probabilities for binary complements that fell significantly short of unity, although analysts' performance was improved through training in Bayesian reasoning using natural sampling trees. In the present research, we tested whether ACH would have a beneficial effect on the degree to which analysts' posterior probability judgments respected the unitarity axiom.

Our investigation into the coherence of analysts' probability judgments was also motivated by the aim of testing the value of statistical, post-judgment methods for improving judgment accuracy. As noted earlier, recent research shows that coherentizing probability judgments so that they respect axioms of probability calculus such as additivity and unitarity can significantly improve judgment accuracy [15]. Moreover, individual differences in the coherence of individuals' judgments can be exploited as a basis for performance weighting contributions to aggregated estimates, making the "crowds wiser" than they would tend to be if each member's contribution had equal weight [44], [45], [46], [47]. Karvetski, Olson, Mandel *et al.* [15] found that the accuracy of probability judgments about the truth of answers to general knowledge questions was improved through coherentizing the judgments, and a further substantial improvement was achieved by coherence weighting the coherentized judgments. In the present research, we examined how effective coherentization and coherence weighting are for improving the accuracy of intelligence analysts' probability judgments. We compared coherentized judgments to raw probability judgments generated with or without the use of ACH. We also compared coherence-weighted aggregate estimates to an equal-weight Linear Opinion Pool (LINOP), which is the arithmetic average across judges [48]. Our interest in this issue was two-fold: First, we aimed to assess the external validity of earlier findings in this nascent area of research on coherentization and coherence-weighted aggregation. Second, we aimed to test whether these post-judgment methods hold promise for organizations, such as intelligence agencies, that generate expert judgment as a product or service.

A further aim of this research anticipated both a possible benefit and possible drawback of ACH. We hypothesized that ACH will not foster greater accuracy in probability judgment because, as we noted earlier, there are processes in the technique, such as disregarding evidential support in information integration, that are normatively indefensible. However, ACH does require analysts to evaluate each piece of information in relation to each hypothesis on the same criterion (consistency). We hypothesized that this might improve analysts' abilities to extract the usefulness of the evidence. Accordingly, we asked analysts to rate the information usefulness of each of the evidential cues presented and we examined how well these ratings correlated, on average, with the probability gain of the cue, a measure of the extent to which knowledge of the cue value is likely to improve classification accuracy [49], [50].

A related aim of ours was to examine whether analysts who display stronger correlations with sampling norms also show better probability judgment accuracy, and whether this "meta-relationship" might differ between ACH and control groups. For instance, ACH proponents might be willing to wager that analysts who use ACH are more likely to reliably encode the information value and to use that information to their advantage by making more accurate judgments.

12.4 METHOD

12.4.1 Participants

Fifty UK intelligence analysts participated in the experiment during regular working hours and did not receive additional compensation for their participation. All participants were pre-registered for intelligence training and were asked by the trainers to participate in the experiment. Mean age was 27.79 years (SD = 5.03) and mean length of experience working as an analyst was 14.08 months (SD = 29.50). Out of 44 participants who indicated their sex, 25 (57%) were male.

12.4.2 Design and Procedure

Participants were randomly assigned in balanced numbers to one of two conditions of the tradecraft factor: the ACH (i.e., tradecraft) condition or the no-ACH (i.e., no tradecraft) control condition. In the ACH condition, participants completed their scheduled ACH training, which was based on Heuer and Pherson [7] and related material from Pherson Associates, LLC. Participants in the control condition received ACH training after the experiment. Participants completed a paper and pencil questionnaire and were subsequently debriefed in small group sessions within the organization in which they worked. However, participants worked individually on the task. Participants in the ACH condition were instructed to approach the judgment task using the eight steps of the ACH method, whereas participants in the control condition were free to use whatever approach they favored. The experiment received ethical approved from the institutional review board of Middlesex University.

12.4.3 Materials

Participants read about a fictitious case in which they were required to assess the tribe membership of a randomly selected person from a region called Zuma.[3] They read that there were four tribes (A-D) that constituted 5%, 20%, 30%, and 45% of Zuma, respectively. Each tribe was then described in terms of 12 probabilistic cue attributes. For instance, for the Acanda tribe (i.e., Tribe A) the description read:

> **Acanda**: 10% of the tribe is under 40 years of age, 75% use social media, 50% speak Zebin (one of two languages spoken in Zuma), 25% are employed, 90% practice a religion, 25% come from a large family (i.e., more than four children), 50% have been educated up to the age of 16, 75% have a reasonably high socio-economic status relative to the general population, 75% speak Zimban (one of

[3] Full instructions for ACH and control conditions are available as supplements.

two languages spoken in Zuma), 75% have a political affiliation, 75% wear traditional clothing, and 25% have fair coloured skin.

Next, the target's cue attributes were described as follows:

The target is under 40 years of age, uses social media, speaks Zebin, is employed, does not practice a religion, does not come from a large family, does not have education up to age 16, does not have a reasonably high socio-economic status, speaks Zimban, is not politically affiliated, wears traditional clothing, and does not have fair coloured skin.

Thus, the target had positive values for half of the cues and negative values for the other half. Furthermore, analysts were told to assume that the target's answers were truthful (due to the administration of a truth serum) in order to ameliorate any possible effects of participants perceiving the information as unreliable or deceptive. Table 12-1 summarizes the informational features of the task.

Table 12-1: Informational Features of Experimental Task. Values represent cue likelihoods.

Evidential cues	Tribe (base rate)				Feature Present in Target
	Acanda (.05)	Bango (.20)	Conda (.30)	Dengo (.45)	
Under 40 years	.10	.10	.90	.90	Yes
Use social media	.75	.50	.25	.50	Yes
Speak Zebin	.50	.75	.50	.25	Yes
Employed	.25	.25	.10	.10	Yes
Practice religion	.90	.90	.10	.10	No
From large family	.25	.50	.75	.50	No
Educated to age 16	.50	.25	.50	.75	No
Have high SES	.75	.75	.90	.90	No
Speak Zimban	.75	.25	.75	.25	Yes
Have political affiliation	.75	.25	.75	.25	No
Wear traditional clothing	.75	.50	.60	.40	Yes
Fair coloured skin	.25	.50	.40	.60	No

In the ACH condition, participants were asked to complete the eight steps of the ACH method (see supplementary materials for full instructions), which included:

a) Identifying all possible hypotheses;

b) Listing significant evidence that is relevant for evaluating the hypotheses;

c) Creating a matrix with all the hypotheses as columns and all items of relevant information as rows and then rating the consistency of each piece of evidence with each hypothesis;

d) Revising the matrix after omitting non-diagnostic evidence;

e) Calculating the inconsistency scores by taking the sum of the inconsistent values and using that to draw tentative conclusions about the relative likelihood of the hypotheses;

f) Analysing the sensitivity of conclusions to a change in the interpretation of a few critical items of relevant information;

g) Reporting conclusions; and

h) Identifying indicators for future observation.

By comparison, in the control condition, participants were asked to "consider the relative likelihood of all of the hypotheses, state which items of information were the most diagnostic, and how compelling a case they make in identifying the most likely hypothesis, and also say why alternative hypotheses were rejected." They were provided with two pages of blank paper on which to respond (none asked for more paper).

All participants completed the same final page of the questionnaire. The first four questions prompted analysts for the probability that the target belonged to each of the four tribes (A-D). Next, they were asked for the probability that the target was friendly and also for the probability that the target was hostile. Probability judgments were made on a 101-point scale that shows numeric probabilities starting at 0 and continuing at every 5% increment up to 100. The instructions noted that 0% meant "impossible" and 100% meant "absolutely certain." Next participants rated on an unnumbered 11-point scale, ranging from not at all to completely, how useful each of the 12 cues was in assessing which the target's tribe membership. For the purpose of statistical analysis, these ratings were entered as values ranging from 1 to 11. We examined analysts' responses to the scale measures of probability and information usefulness.

12.4.4 Coherentization and Coherence Weighting

As described previously, more often than not, individuals produce probability estimates that are incoherent and violate probability axioms, and there is evidence that more coherent estimates are associated with more accurate estimates [51]. Given a set or vector of elicited probabilities that is incoherent, the Coherent Approximation Principle (CAP) [44], [45], [52] was proposed to obtain a coherent set of probabilities that is minimally different in terms of Euclidean distance from the elicited probabilities with the goal of improving accuracy. This "closest" set of coherent probabilities is found by projecting the incoherent probabilities onto the coherent space of probabilities. An incoherence metric can then be defined as the Euclidean distance from an incoherent set of probabilities to the closest coherent set of probabilities. For example, if an analyst in the present research provided probability judgments of .2, .3, .4, and .3 for the four MECE hypotheses A-D, respectively, these estimates are incoherent because they sum to a value greater than 1 and thus violate the unitarity constraint. Using the CAP and (if needed) quadratic programming [15] a coherent set of recalibrated probabilities can be obtained, which minimizes the Euclidean distance between the point {.2, .3, .4, .3} and all quartet vectors with values between 0 and 1, such that the sum of the four values is 1. For this example, the probabilities of .15, .25, .35, and .25 represent the closest coherent set, with minimum distance as follows:

$$\sqrt{(.2-.15)^2+(.3-.25)^2+(.4-.35)^2+(.3-.25)^2} = .10$$

The resulting value, moreover, represents an Incoherence Metric (IM), expressed, more generally, as:

$$\mathrm{IM} = \sqrt{\sum_{i=1}^{k}(y_i - y_i^c)^2}. \tag{12-1}$$

In Equation (12-1), IM is calculated over the sum of k judgments that form a related set, and notably IM is zero when the provided judgments are perfectly coherent. The CAP is not limited to using only the unitarity constraint but can be applied with any set of coherence constraints that can be defined mathematically as an optimization program.

As noted earlier, variations in IM across individuals can also be used as a basis for performance-weighted aggregation. With IM_j as the incoherence metric for the j^{th} individual in an aggregate, a weighting function should satisfy general properties. First, it should be strictly decreasing as IM_j increases, thus assigning harsher penalties to more incoherent individuals. Because weights are normalized during the aggregation, only the ratio values of weights are relevant. Thus, the function can be arbitrarily scaled in the 0 – 1 interval, with 1 representing a perfectly coherent judge. In the present research, we use a weighting function similar to that of Wang *et al.* [47].

$$\omega_j = e^{(-IM_j \cdot \beta)}. \tag{12-2}$$

The weighting function assigns full weight to the j^{th} individual if $IM_j = 0$ or if $\beta = 0$. In the former case, this is due to the perfect coherence of j's raw estimates, while in the latter case the weighting function is non-discriminatory and equivalent to taking the arithmetic average across individuals.

Next, we define the coherence-weighted average of n (where $2 \leq n \leq N$) individuals' coherentized probability judgments of the i^{th} hypothesis as:

$$\bar{y}_i^{cc} = \sum_{j=1}^{n} \omega_j y_{ij}^c \Big/ \sum_{j=1}^{n} \omega_j . \tag{12-3}$$

Again, if $\beta = 0$, we have an equal-weighted (arithmetic) average of the coherentized judgments:

$$\bar{y}_i^c = \frac{1}{n} \sum_{j=1}^{n} y_{ij}^c . \tag{12-4}$$

Note that the coherence constraints on y_{ij}^c imply that set of all coherent probabilities is a convex set, and any linear combination of elements from a convex set is again an element of the same set. Therefore, the aggregated estimates must also be coherent and do not have to be coherentized again.

In the present research, we let $\beta = 5$, but we later show that the results are not sensitive to the exact value chosen. Choosing a sufficiently large value alleviates the issue with the "fifty-fifty blip", which results when an individual expresses epistemic uncertainty by responding .5 over multiple judgments [52]. In the present research, if an analyst entered 0.5 for each hypothesis, *A-D*, the values would sum to 2, and the participant's IM score would be .50. In the weighting function, we have $\omega(.50) = .082$. This participant would be assigned only 8.2% of the weight that would be assigned to a perfectly coherent participant.

12.4.5 Metrics

The primary measure of accuracy we use is Mean Absolute Error (*MAE*), which in this research computes the mean absolute difference between a human-originated judgment (i.e., raw, transformed, or aggregated), y_i , and the corresponding posterior probabilities derived from Bayes theorem assuming class conditional independence (i.e., a "naïve Bayes" model), x_i . We acknowledge that this simplifying assumption is not necessitated by the task. However, we believe it is reasonable to assume that participants did not perceive conditional dependence and subsequently take it into account – at least we found no evidence to support such a conclusion in participants' responses. Using the naïve Bayes model, xA = .08, xB = .15, xC4 = .46, and xD = .31.

Accordingly,

$$\mathrm{MAE} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n}|y_{ij} - x_i|. \tag{12-5}$$

The summation over *i* refers to the set of hypotheses (i.e., in this research, k = 4).

An advantage of *MAE* over mean squared error or root mean squared error is that it is less susceptible to outliers [53], [54]. In addition, *MAE* is decomposable into quantity disagreement (*QD*) and allocation disagreement (*AD*):

$$\mathrm{QD} = |\mathrm{ME}|, \text{ where } \mathrm{ME} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - x_i). \tag{12-6}$$

$$\mathrm{AD} = \mathrm{MAE} - \mathrm{QD}. \tag{12-7}$$

QD is the absolute value of Mean Error (*ME*) or bias. *AD* represents remaining inaccuracy after removal of *QD* (i.e., absolute bias), which necessarily involves a fair balance between underestimations and overestimations of correct values (i.e., any imbalance is part of *QD*). Coherentization reduces *MAE* by eliminating *QD*.

As noted earlier, we used a measure of classification accuracy improvement called probability gain [50] to assess analysts' accuracy in rating cue usefulness:

$$\text{probability gain}(Q) = [\sum_{q_j} P(q_j) \times \max_{h_i} P(h_i \mid q_j)] - \max_{h_i} P(h_i). \tag{12-8}$$

12.5 RESULTS

12.5.1 Coherence of Probability Judgments

We tested the coherence of analysts' probability judgments as a function of tradecraft using the following logical constraints:

$$y_A + y_B + y_C + y_D = 1 \text{ (unitarity, quarternary partition).} \tag{12-9}$$

$$y_H + y_F = 1 \text{ (unitarity, binary partition).} \tag{12-10}$$

$$y_H = y_A + y_C,\ y_F = y_B + y_D \text{ (additivity, two binary partitions).} \tag{12-11}$$

Equations (12-9) and (12-10) reflect the unitarity axiom and Equation (12-11) reflects the additivity axiom.[4] In violation of Equation (12-9) and showing a strong unpacking effect, the sum of the probabilities assigned to the four MECE hypotheses significantly exceeded unity in the control condition (M = 1.54 [1.33,1.76], t[24] = 13.63, d = 0.96, p < 001) and in the ACH condition (M = 1.77 [1.56, 1.97], t[24] = 19.53, d = 1.69, p < .001). The unpacking effect did not significantly differ between conditions, but the difference in the size of

[4] Square brackets show bootstrapped 95% confidence intervals from 1,000 bias-corrected and accelerated samples, Δ denotes the mean difference between conditions, and *d* refers to the effect size estimator, Cohen's *d*.

these effects was nevertheless of medium effect size by Cohen's [55] standards and favored the control group, = 0.22[-0.08, 0.55], *t*[45.8)] = 1.55, d = 0.45, p = .13. In contrast, but consistent with several studies also finding unitarity for binary complements [33], [36], [37], [56], the total probability assigned to the binary complements, H and F, did not significantly differ from unity in either the control condition (M = 0.98 [0.90,1.04], *t*[24] = 27.38, d = 0.12, $p < .001$) or the ACH condition (M = 0.95 [0.83, 1.00], *t*[24] = 23.45, d = 0.25, p < .001). Thus, on average, analysts respected the unitarity constraint imposed by Equation 12-10.

Turning to tests of additivity, we computed the Sum of the (Signed) Non-additivity (SSN):

$$SSN = (y_A + y_C - y_H) + (y_B + y_D - y_F). \quad (12\text{-}12)$$

If Equation 12-12 is respected, *SSN* = 0. However, it is evident that implicit disjunctions were assigned significantly less probability than what was assigned, in sum, to their constituents in both the ACH condition (*M* = 0.82 [0.64, 0.99], *t*[24] = 8.98, *d* = 1.80, $p < .001$) and the control condition (*M* = 0.56 [0.37, 0.74], *t*[24] = 5.34, *d* = 1.07, $p < .001$). In addition, mean additivity violation, consistent with the unpacking effect, was marginally greater in the ACH condition than in the control condition, = 0.25 [−0.05, 0.57], *t*(48) = 1.81, *d* = 0.52, *p* = .08. Once again, this difference was of medium effect size.

12.5.2 Accuracy of Probability Judgments

As noted earlier, we compared the accuracy of analysts' untransformed (i.e., not coherentized) probability judgments for the four-way MECE partition (i.e., Tribes *A-D*) using analysts' *MAE* calculated over the four estimates. Although there was a significant degree of inaccuracy in both the control condition (*MAE* = 0.21 [0.17, 0.26], *t*[24] = 9.69, *d* = 1.94, $p < .001$) and the ACH condition (*MAE* = 0.26 [0.22, 0.29], *t*[24] = 14.39, *d* = 2.88, $p < .001$), the effect of tradecraft was not significant, = 0.04 [-0.02, 0.11], *t*(45.9) = 1.49, *d* = 0.43, *p* = .14. Nevertheless, as the effect-size estimate reveals, there was a medium-sized effect of tradecraft that, once again, favored the control group.

The observed *MAE* in the sample was also compared to that obtained from 10,000 random draws of probability values for each of the four hypotheses, A-D (i.e., where each probability was drawn from a uniform distribution over the [0, 1] interval – a simulated dart-throwing chimp, to use Tetlock's [57] metaphor). *MAE* for the random judgments was 0.33. Thus, analysts performed significantly better than chance, analysts' *MAE* = 0.23 [0.21, 0.26], t(49) = 6.69, d = 0.95, p < .001.

Given that the QD decomposition of *MAE* calculated over the four MECE hypotheses is directly related to unitarity violation and, further, given that we have established that this type of coherence violation is greater in the ACH condition than in the control condition, we can verify that the proportion of total inaccuracy (*MAE*) accounted for by QD is greater in the ACH condition than in the control condition. In fact, this was confirmed: The QD/*MAE* proportion was .73 [.60, .86] in the ACH condition and .50 [.32, .67] in the control condition, a significant effect of medium size, = .23 [.04, .45], t(45.7) = 2.08, d = 0.60, p = .04.

Although the preceding analyses do not indicate that ACH helps to improve analysts' probability judgments, critics might argue that the method is not aimed at minimizing absolute error but rather at improving the rank ordering of alternative hypotheses in terms of their probability of being correct. To address this point, we calculated the rank order (Spearman) correlation between each analyst's four raw probability judgments of A-D and the probability vector of the naïve Bayes model. The mean correlations in the ACH condition (M = .29 [.02, .55]) and the control condition (M = .24 [-.08, .55] did not significantly differ, *t*[46.9] = 0.28, d = 0.08, p = .78. Therefore, we find no support for the hypothesis that ACH helped analysts to better assess the relative probability of the four hypotheses.

12.5.3 Information Usefulness

As noted earlier, we hypothesized that the consistency rating process in ACH, which requires analysts to assess each piece of evidence for consistency with each hypothesis, and the subsequent diagnosticity assessment process, which requires analysts to consider information usefulness, might help analysts capture variation in information utility. Accordingly, we computed the Pearson correlation between each analyst's ratings of the information usefulness of the 12 cues and the probability gain values for those cues. Providing support for the preceding hypothesis, the mean correlation in the ACH condition (M = .68 [.61, .75]) was significantly greater than the mean value in the control condition (M = .17 [−.02, .35]), and the effect size was very large, $t[29.5] = 5.35$, d = 1.59, p < .001.

Next, we examined whether these correlations were themselves related to analysts' *MAE* scores. Overall, this correlation was non-significant, r (49) = -.14 [-.39, .15], p = .53. However, the observed relationship was strikingly different between the two conditions. The correlation was negligible in the ACH condition, r(24) = -.10 [-.44, .28], p = .63,but it was significant and of medium-to-large effect size in the control condition, r(24) = -.42 [-.69, -.07], p = .045. Although analysts using ACH were more likely than analysts in the control condition to track the variation in probability gain with their usefulness ratings, the degree to which the ACH group tracked probability gain had almost no correspondence to their accuracy, whereas it did for the control group.

12.5.4 Recalibrating Probability Judgments

The substantial degree of non-additivity observed in analysts' probability judgments implies that recalibration procedures that coherentize the judgments will not only ensure coherence; they will also benefit accuracy by eliminating the QD component of *MAE*. Thus, we coherentized analysts' probability judgments of A-D so that they respected the unitarity constraint in Equation (12-9).[5] The coherentized probability judgments (*MAE* = 0.15 [0.13, 0.18]) were significantly more accurate than the raw judgments (*MAE* = 0.23 [0.21, 0.26]), = -0.08 [-0.11, -0.06], $t[49] = 6.77$, d = 0.96,p < .001. This represents a 35% reduction in *MAE* and approximately a 1 SD improvement. Recall that the proportion of *MAE* attributable to QD was significantly greater in the ACH condition than in the control condition. This suggests that the effect of coherentizing will be stronger in the ACH condition. In fact, d = 0.69 in the control condition and d = 1.37 in the ACH condition. Therefore, the SD improvement is roughly twice as large in the ACH condition as it is in the control condition. Moreover, after coherentizing, the effect of tradecraft on accuracy is negligible, = 0.01 [-0.04, 0.07].

We once again compared analysts' judgment accuracy to the performance of the average dart-throwing chimp. However, this time we coherentized the randomly generated probabilities, which yielded *MAE* = 0.21, a value that was significantly inferior to the observed coherentized *MAE* of 0.15, $t[49] = 4.56$, d = 0.64, p < .001. An alternative method of assessing chance is to define it in terms of all possible permutations of the probabilities actually provided by each participant, rather than as a uniform distribution. Using this definition, the superiority of the analysts over chance was still apparent but not as large: 0.16 for chance, 0.13 for the participants ($t(49) = 2.75$, p = .008), a difference of about 0.03 rather than 0.06.[6] This analysis suggests that probability judgments were in the right range, but they were conveying very little information about the relative probabilities of the four hypotheses, this reducing the power of the experiment to detect group differences.

[5] An alternative form of coherentization that used Equations (12-9) and (12-11) was also tested but found to be virtually indistinguishable. Thus, we used the simpler form.

[6] Jon Baron conducted this analysis and used normalization (i.e., dividing each stated probability by the sum of the four) rather than coherentization as the recalibration method.

12.5.5 Aggregating Probability Judgments

Coherentization yielded a large improvement in the accuracy of analysts' probability judgments. We examined how much further improvement in accuracy might be achieved by aggregating analysts' probability judgments. To do so, we generated 1,000 bootstrap samples of statistical group sizes ranging from 1 (i.e., no aggregation) to 49 in increments of two. We aggregated probability judgments in two ways: using an unweighted arithmetic average of coherentized probability judgments and using a coherence-weighted average of such estimates.[7] We examined the effect of aggregation on *MAE* as well as on the average Spearman correlation between the aggregated estimates and the vector of values from the naïve Bayes model. As a benchmark, we also examined the effect of these aggregation methods on random responses, where each data point is based on 1,000 simulations of probability judgments from a uniform distribution over the [0, 1] interval.

As Figure 12-1 shows, these analyses yield several important findings. First, they confirm that, when aggregated, analysts' judgments are substantially more accurate than aggregated random judgments. Second, it is evident from the left panel in Figure 12-1 that aggregation greatly improves accuracy in analysts' judgments, but to a degree comparable to that observed in the randomly generated response data. This suggests that most of the error reduction observed is due to variance reduction from averaging and should not be attributed to an eking out of any crowd wisdom, as clearly there is no wisdom in the random response data.[8] Third, it is equally evident from the right panel in Figure 12-1 that aggregation over increasingly larger group sizes steadily increases the correct rank ordering of probabilities. This effect is clearly not manifested in random response data, where aggregation has no benefit. Fourth, aggregation with coherence weighting did not outperform aggregation with equal weighting; in fact, it slightly underperformed. Finally, the left panel in Figure 12-1 shows that most error reduction due to aggregation was achieved with small group sizes.

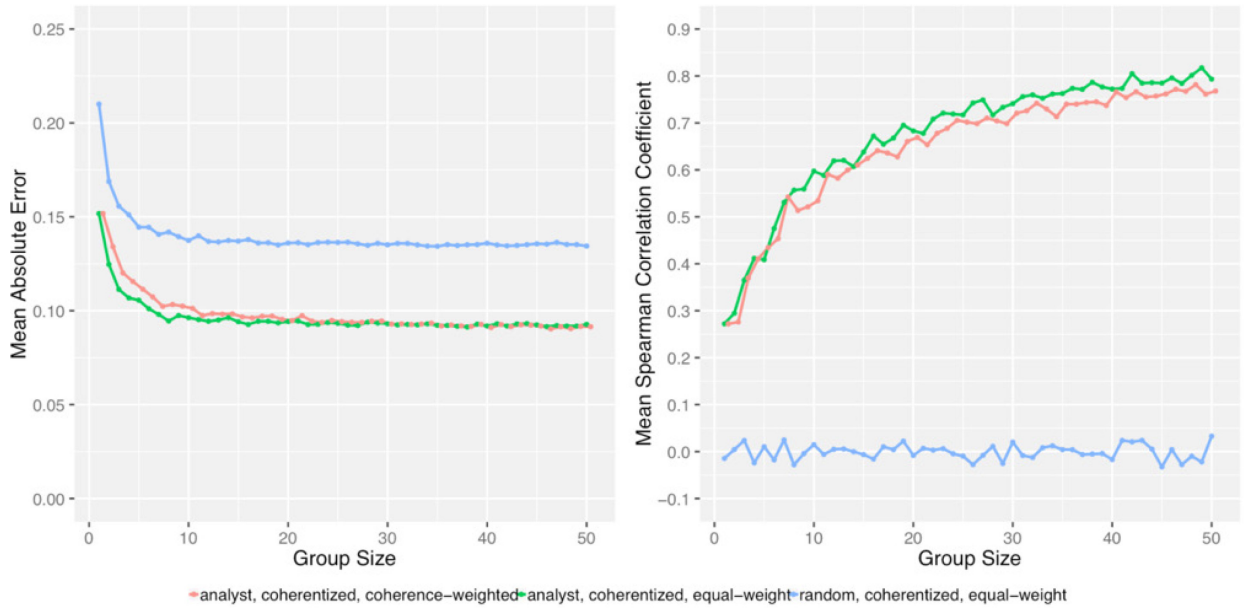


Figure 12-1: Accuracy of Probability Judgments by Group Size and Aggregation Method.

[7] For coherence-weighted aggregation, β = 5." However, as shown in the supplementary figure, the effect of coherence weighting was robust across a wide parametric range.

[8] Note that aggregation of random responses will bring all responses closer to .25. In the limit, the *MAE* of this constant response may be lower than that for a set of responses with excessive variability.

Figure 12-2 clarifies that there was a significantly greater the proportion of cases where *MAE* was lower for a group size of two than for single individuals (i.e., the probability of improvement), and likewise the stepwise increase in group size from two to three significantly increased the proportion with lower *MAE* scores. However, no additional stepwise increase in group size yielded significant improvements.

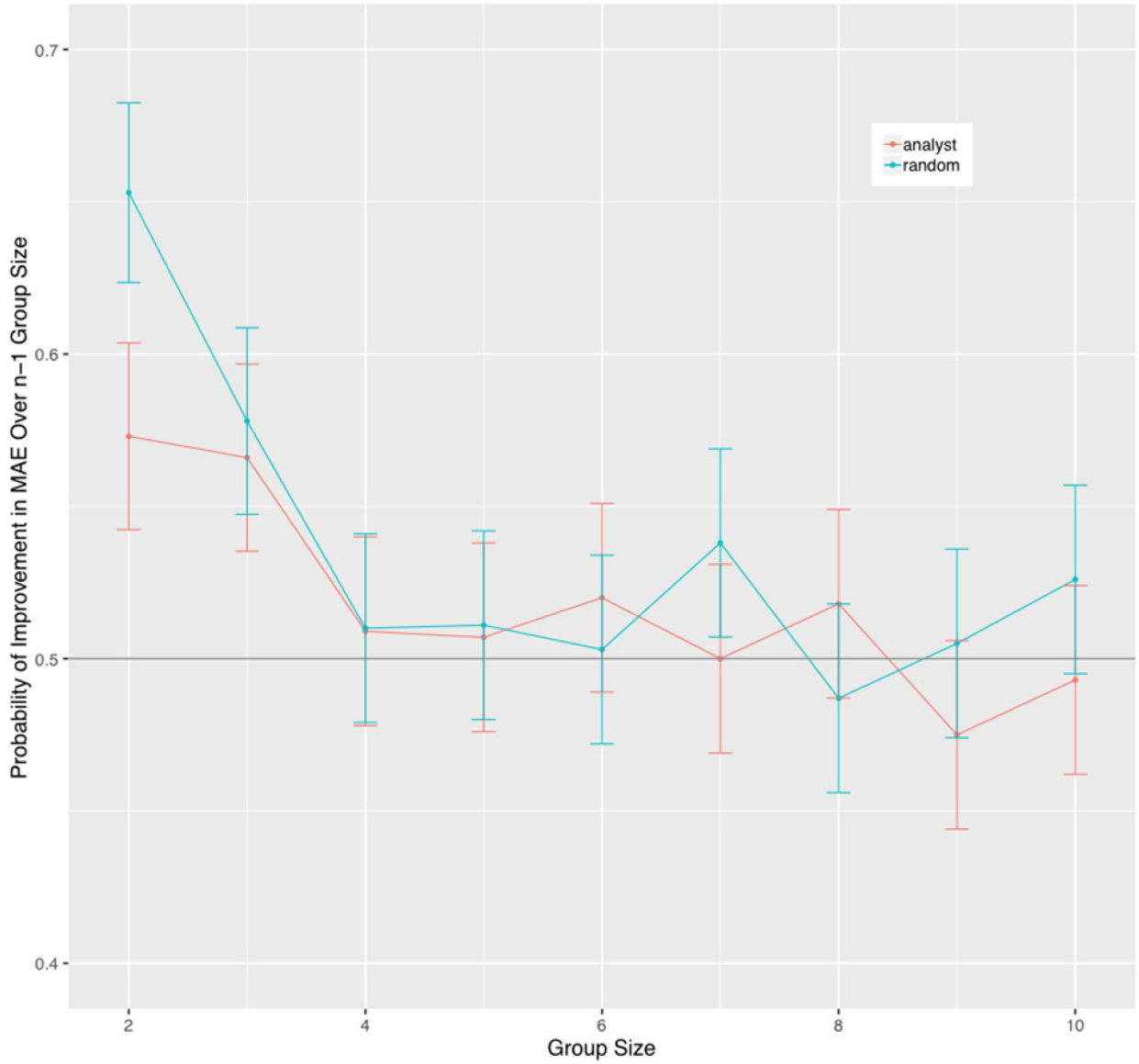


Figure 12-2: Probability of Improvement Achieved by Increasing Group Size by One Member. Bars show 95% confidence intervals from 1,000 bootstrap samples. The reference line shows the probability of improvement by chance.

Finally, we assessed the proportional gain in accuracy achieved by recalibration and equal-weight aggregation when n = 50. As noted earlier, coherentizing the disaggregated judgments yielded a 35% reduction in *MAE*. If we combine coherentization with equal-weight aggregation of the full sample of 50 analysts, we obtain *MAE* = 0.09, a 61% reduction in *MAE* over the value for analysts' original probability judgments (i.e., *MAE* = .23). That is, 61% of the inaccuracy of analysts' probabilistic assessments of the target's category membership was eliminated by first coherentizing those assessments and then taking an unweighted average of them prior to scoring.

12.6 DISCUSSION

Although intelligence organizations routinely train and advise analysts to use tradecraft methods, such as ACH, to mitigate cognitive biases and thereby improve the coherence and accuracy of their assessments, there has been a dire lack of research on their effectiveness. The present research conducted such a test and found that ACH failed to improve intelligence analysts' probabilistic judgments about alternative hypotheses. It even had a small detrimental effect on some measures of coherence and accuracy. In such cases, the comparison between conditions yielded a medium effect size in favor of the control group. To better understand the advantage of not using ACH in the present research task, it is helpful to convert the effect size into a stochastic superiority or probability of superiority estimate equal to the area under the receiver-operator characteristic curve in signal detection theory [58], [59], [60]. The probability of superiority is the probability that a randomly selected member of one condition will outperform a randomly selected member of another condition. For accuracy, for instance, the effect size, d = 0.45, yields a probability of superiority estimate equal to .62 favoring the control condition. That is, if one analyst were randomly drawn from the ACH condition and another randomly drawn from the control condition, there would be a 62% chance of the former having worse accuracy than the latter.

Comparable probabilities of superiority favoring the control condition likewise are obtained in tests of unitarity and additivity. In each case, coherence violations conformed to the unpacking effect predicted by support theory [32], [33]. As noted earlier, the unpacking effect refers to the tendency for people to assign greater total probability to the sum of a MECE partition of a disjunctive event (in the case of additivity) or an event space (in the case of unitarity). The unpacking effect has been shown to undermine the logical coherence of geopolitical assessments [61], which suggests that such forms of incoherence can undermine strategic intelligence assessments. Indeed, compared to regular forecasters, elite superforecasters of geopolitical topics tend to display greater coherence on other, unrelated probabilistic tasks [51]. It should concern the intelligence community that a commonplace analytic tradecraft technique served to increase (rather than reduce) this form of judgmental error.

Of course, critics might argue that perhaps analysts interpreted the request for probabilities as requiring only a relative probability assessment of the hypotheses. After all, ACH is primarily aimed at ranking hypotheses by likelihood, and for that reason our control analysts were also instructed to assess the relative likelihoods of the hypotheses. However, we elicited probabilities for each hypothesis separately on a scale covering the [0, 1] interval. Moreover, the four probabilities (A-D) that are bound by the unitarity axiom were elicited in immediate succession, an elicitation feature shown to mitigate incoherence [36]. Therefore, we expect to find even greater incoherence in analytic practice where the logical relations between assessments are likely to be obscured. Finally, we found that the rank-order correlations between analysts' judgments and the correct values were small, on average, having only about 7% shared variance.

Another striking result of the present research concerns the relationship between the quality of analysts' information usefulness evaluations and the quality of their probability judgments regarding the alternative hypotheses. Although analysts who used ACH provided ratings of probabilistic cue usefulness that were more strongly correlated with the cues' probability gain values than analysts who did not use ACH, the former group's assessments of information usefulness did virtually nothing to guide them to exploit the knowledge effectively to boost accuracy in probability judgments. In contrast, among analysts in the control group, there was substantially better correspondence between accuracy and the degree to which their usefulness ratings tracked probability gain. Analysts in the control group whose usefulness ratings tracked probability gain were better poised than analysts in the ACH group to use that knowledge to improve the accuracy of their probability assessments. This finding was unanticipated and should ideally be tested for reproducibility in future research.

While speculative, one explanation for the disconnect between accurate evaluation of information usefulness and accuracy of probability judgments is that the consistency encoding phase in ACH prompts analysts to

adopt a perspective that is evidence contingent rather than hypothesis contingent. That is, analysts are taught to evaluate evidence hypothesis consistency within pieces of evidence and across hypotheses rather than the other way around. This approach is deliberate, reflecting Heuer's [18] belief that analysts are susceptible to confirmation bias and thus need to be made to focus on evidence rather than their preferred hypothesis. The evidence contingent approach should prompt consideration of information usefulness given that the consistency between a piece of evidence and each hypothesis being evaluated is assessed before proceeding to another piece of evidence. However, we see that information integration within hypotheses is left to the questionable "sum of the inconsistency scores" rule in ACH. Unlike a normative (e.g., Bayesian) approach, this rule merely serves as a summator and, moreover, selectively so by choosing to ignore scores that indicate degree of positive consistency. The integration rule is also exceptionally coarse in its treatment of evidence, assigning one of only three levels (-2, -1, 0) to each piece of evidence, and such coarseness is likely to impede judgment accuracy [62].

Moreover, ACH does virtually nothing as an analytic support tool to ensure that analysts consistently map evidential strength onto -1 and -2 ratings. Consider two hypotheses, A and B. Assume that given five pieces of evidence, three analysts, X, Y, and Z agreed on the following. All five pieces of evidence are inconsistent with A and three pieces are inconsistent with B. Assume further that compared to Y, X has a low threshold for assigning -2 ratings, and Z has a high threshold. All three analysts might agree that the five pieces of evidence are inconsistent with A, but not strongly so, and they would assign -1 for each piece. They might further agree that the three pieces of evidence that are inconsistent with B are stronger in their inconsistency than in the case of A, but given their differing thresholds for assigning -2 ratings, they may vary in their ratings. For instance, X might assign -2 to the three pieces that are inconsistent with B, Y might assign, two -2 ratings and one -1 rating, and Z might assign -1 ratings to each of the three pieces of evidence inconsistent with B. If so, in spite of the substantial agreement among analysts, using ACH, X would judge A less probable than B, Y would judge A and B as equally probable, and Z would judge A as more probable than B!

The present findings indicate that ACH is ineffective as a means of supporting analysts in assessment tasks requiring the integration of uncertain evidence in order to evaluate a set of hypotheses. The findings challenge a widespread assumption among tradecraft professionals in intelligence organizations that, although ACH (and SATs, in general) might not always help the analyst, at least they don't hurt the analyst [11]. Two of the authors (DRM and MKD) who have worked for several years with analytic tradecraft professionals have repeatedly encountered a "nothing to lose" attitude when it comes to SAT training and on-the-job use. Yet, our findings suggest that, in fact, ACH can impede the quality of intelligence assessments. It can do so in two ways: first, by undermining the coherence and accuracy of estimates and, second, by fostering a disconnection between evidence evaluation and hypothesis evaluation. We therefore urge intelligence organizations to be more circumspect about the benefits of training analysts to use ACH and other SATs that have not received adequate testing.

Indeed, a commonplace rebuttal from intelligence professionals to any criticism of tradecraft methods is that although they aren't perfect, intelligence organizations can't just "do nothing." The ideas of leaving analysts to their own "intuitive" reasoning is thought to – and often does – result in bias and error. Our findings challenge this assumption since analysts who were left to their own devices performed better than analysts who used ACH.

SAT proponents are likely to object and claim that our findings lack external validity. After all, intelligence analysts seldom are presented with such neat problems where all evidence is precisely quantified and expresses relative frequencies and where the full set of pertinent hypotheses is explicit and, further, it is evident that these hypotheses are also neatly partitioned (i.e., MECE). We agree that in these and other respects the experimental task we used lacks mundane realism. However, we disagree with the implications that proponents would likely draw from such observations. Intelligence problems are murkier in many respects – the quality of evidence will be variable, the hypotheses might be unclearly defined and will often fail to yield a MECE set, and analysts are likely to give no more than vague probability estimates on coarse

verbal probability scales [8], [22], [62], [63]. We see no compelling reason why ACH should help under those conditions when it does not help hypothesis evaluation under the much more modest requirements of the present experimental task. Indeed, it is possible that ACH can do even more harm to judgment when analysts use it on the job.

Clearly, it would be beneficial to conduct research in the future that uses tasks that are more challenging in the respects noted while permitting unambiguous evaluation of the merits of ACH or other SATs. However, the present research already shows that ACH is not an all-purpose judgment corrective for problems involving the evaluation of multiple hypotheses on the basis of uncertain evidence. In fact, the poor performance of both groups of analysts in this research raises a more basic question: why were they so inaccurate on a task (even in terms of their relative probability judgments) that is arguably much easier than the types of so-called puzzles and mysteries they encounter on the job? This may ultimately prove to be a more important finding than the relative performance between conditions. In the present task, analysts had unambiguous sources of accurate information that they could exploit, yet most were at a loss do so regardless of whether they used ACH or not. Our findings therefore raise a fundamental question about the competence of analysts to judge probabilities. Given the small and homogeneous sample of analysts we tested, it would be wrong to draw sweeping generalizations. Yet, if our findings do generalize across a wide range of analyst samples, it should prompt the intelligence community and the bodies that provide intelligence oversight to take stock of the practical significance of the findings and study the putative causes of poor performance.

We also respond to SAT proponents by noting that "doing nothing" is not the only alternative to using conventional analytic tradecraft techniques such as ACH. In this research, we examined two promising statistical methods that intelligence organizations could use to improve probability judgments after analysts had provided judgments – methods we accordingly describe as post-analytic. One method, coherentization, exploits the logical structure of related queries by recalibrating probability assessments so that they conform to one or more axioms of probability calculus. As noted earlier, Karvetski, Olson, Mandel *et al.* [15] showed that such methods substantially improve the accuracy of probability judgments. Likewise, in the present experiment, a large improvement in analysts' accuracy was achieved by coherentizing analysts' probability judgments such that they respected the unitarity axiom. This method fully counteracted the unpacking effect exhibited by analysts in this research, especially those who were instructed to use ACH. We view CAP-based coherentization as illustrative rather than definitive. Other recalibration methods might be even more effective or easier to apply. For instance, in the present research, we could have coherentized probabilities by simple normalization (i.e., dividing each by their sum), as researchers sometimes do as a step in the statistical analysis of probability judgment data [64]. This method would have yielded even slightly better accuracy than CAP-based coherentization (*MAE* = 0.13 for normalization, vs. 0.15 for CAP). Our study is clearly not designed to examine such competitions given it relies on a single vector of values defining probabilistic accuracy. However, our findings suggest that research comparing optimization methods using such techniques under a broad range of task conditions are needed.

Another post-analytic method that intelligence organizations could use to boost the accuracy of probabilistic assessments is to aggregate them across small numbers of analysts. We found that substantial benefits to accuracy were achieved by taking the arithmetic average of as few as three analysts. These findings are consistent with earlier studies showing that most of the advantage from aggregating can be achieved with between two to five judges [65], [66], [67]. Moreover, we found that a simple equal-weighted aggregate of analysts' judgments yielded comparable benefit to the more complex coherence-weighted aggregation method. This result was unexpected given the superior performance coherence weighting afforded over equal weighting in recent studies [15], [47]. A key difference between the tasks in the present research and Karvetski, Olson, Mandel *et al.* [15] is that the former included all information relevant to solving the task, whereas the latter relied on participants' knowledge of world facts, such as who was the first person to walk on the moon. Thus, whereas in the present research, coherentization may have already reaped most of the benefit achievable through coherence weighting, in the earlier studies coherence weighting might also have benefited accuracy by predicting how knowledgeable participants were.

More generally, the present results indicate that intelligence organizations should be exploring how to effectively incorporate processes for eliciting judgments from multiple analysts and then aggregating them in order to reduce judgment error. At present, intelligence organizations rarely capitalize on statistical methods such as the recalibration and aggregation approaches shown to be effective in the present research. Instead, the management of intelligence production tends to rely on traditional methods such as having sole source analysts provide input to an all-source analyst (an approach that is common at the operational level), or by having a draft intelligence report reviewed by peers with relevant domain expertise and by the analyst's director (an approach often employed at the strategic level). Still, we caution not to infer too much from the aggregation results. It is tempting to suggest that the aggregate divines the wisdom of crowds, as Surowieki [68] put it, yet our finding that aggregation of random response data yielded comparable error reduction as in analysts' judgments clearly challenges that interpretation as there was no wisdom in the random data to divine. Our analysis of how aggregates can improve relative probability assessment, however, showed a large improvement inaccurately capturing the rank ordering of probabilities, and this benefit was entirely absent in the random response data, which suggests that aggregation did in fact boost the signal-to-noise ratio in analysts' ordered probability judgments.

To conclude, we argue that the intelligence community should look to recent examples of research that illustrate how organizations could better integrate recalibration and aggregation methods pioneered in decision science into day-to-day analytic practices. One example involves the systematic monitoring of probabilistic forecast accuracy within intelligence organizations [63], [69]. The results of such monitoring have shown that analysts' forecasts tend to be underconfident, and that the calibration of intelligence units can be improved post-judgment through an organizational recalibration process that "extremizes" overly cautious forecasts [12], [13], [14]. Another example is the introduction in the US intelligence community of a classified prediction market that poses forecasting questions not unlike those worked on by strategic analysts as part of their routine assessment responsibilities. Stastny and Lehner [70] showed that analysts' forecasts within the prediction market, which aggregated the forecasters' estimates but also shared the aggregated estimates with the forecasters, were substantially more accurate than the same forecasts arrived at through conventional analytic means. These examples illustrate the benefits to analytic accuracy and accountability that intelligence organizations could accrue if they leveraged post-analytic mathematical methods for boosting the quality of expert judgment.

12.7 REFERENCES

[1] Belton, I., and Dhami, M.K. (in press). Cognitive biases and debiasing in intelligence analysis. In: *Handbook on Bounded Rationality*, Viale, R., and Katzikopoulos, K. (Eds.). London, UK: Routledge.

[2] Chang, W., Berdini, E., Mandel, D.R., and Tetlock, P.E. (2018). Restructuring structured analytic techniques in intelligence. *Intelligence and National Security,* 33(3):337-356.

[3] Coulthart, S.J. (2017). An evidence-based evaluation of 12 core structured analytic techniques. *International Journal of Intelligence and CounterIntelligence,* 30(2):368-391.

[4] Marchio, J. (2014). Analytic tradecraft and the intelligence community: Enduring value, intermittent emphasis. *Intelligence and National Security,* 29(2):159-183.

[5] Butler, F.E.R., Chilcot, J., Inge, P.A., Mates, M., and Taylor, A. (2004). *Review of Intelligence on Weapons of Mass Destruction,* the Butler Review, HC 898. London, UK. Retrieved from http://news.bbc.co.uk/nol/shared/bsp/hi/pdfs/14_07_04_butler.pdf

[6] Dhami, M.K., Belton, I.K., and Careless, K.E. (2016). Critical review of analytic techniques. In: *Proceedings of the 2016 European Intelligence and Security Informatics Conference*, 152-155.

[7] Heuer, R.J., Jr., and Pherson, R.H. (2014). *Structured Analytic Techniques for Intelligence Analysis*. Developed by Pherson Associates, LLC.Washington DC: CQ Press.

[8] Dhami, M.K., Mandel, D.R., Mellers, B.A., and Tetlock, P.E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science,* 10(6):753-757.

[9] National Research Council. (2011). *Intelligence Analysis for Tomorrow: Advances from the Behavioral and Social Sciences*. Washington DC: National Academies Press.

[10] Pool, R. (2010). *Field Evaluation in the Intelligence and Counterintelligence Context: Workshop Summary*. Washington DC: The National Academies Press.

[11] Mandel, D.R., and Tetlock, P.E. (2018). Correcting judgment correctives in national security intelligence. *Frontiers in Psychology*, 9:2640.

[12] Mandel, D.R., and Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences of the United States of America*, 111(30):10984-10989.

[13] Baron, J., Mellers, B.A., Tetlock, P.E., Stone, E., and Ungar, L.H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis,* 11(2):133-145.

[14] Turner, B.M., Steyvers, M., Merkle, E.C., Budescu, D.V., and Wallsten, T.S. (2014). Forecast aggregation via recalibration. *Machine* Learning, 95(3):261-289.

[15] Karvetski, C.W., Olson, K.C., Mandel, D.R., and Twardy, C.R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis,* 10:305-326.

[16] Mandel, D.R., Karvetski, C., and Dhami, M.K. (2018). Boosting intelligence analysts' judgment accuracy: What works, what fails? *Judgment and Decision Making,* 13 (6):607-621.

[17] Canadian Safety and Security Program Projects CSSP-2016-TI-2224 (*Improving Intelligence Assessment Processes with Decision Science*) and CSSP-2018-TI-2394 (*Decision Science for Superior Intelligence Production*), Department of National Defence Project 05da (*Joint Intelligence Collection and Analytic Capability*), and HM Government.

[18] Heuer, R.J., Jr. (1999). *The Psychology of Intelligence Analysis*. Washington, DC: Central Intelligence Agency, Center for the Study of Intelligence.

[19] US Government. (2009). *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis*. Washington, DC: Center for the Study of Intelligence Analysis.

[20] UK Ministry of Defence. (2013). *Quick Wins for Busy Analysts*. London, UK: UK Ministry of Defence.

[21] Popper, K. (1959). *The Logic of Scientific Discovery*. London, UK: Hutchison & Co.

[22] Mandel, D.R. (2019). Can decision science improve intelligence analysis? In: *Researching National Security Intelligence: Multidisciplinary Approaches,* Coulthart, S., Landon-Murray, M., and Van Puyvelde, D. (Eds.). 117-140. Washington DC: Georgetown University Press.

[23] Jones, N. (2018). Critical epistemology for analysis of competing hypotheses. *Intelligence and National Security*, 33(2):273-289.

[24] Karvetski, C.W., Olson, K.C., Gantz, D.T., and Cross, G.A. (2013). Structuring and analyzing competing hypotheses with Bayesian networks for intelligence analysis. *EURO Journal on Decision Processes,* 1(3-4):205-231.

[25] Pope, S., and Jøsang, A. (2005). Analysis of competing hypotheses using subjective logic. In: *Proceedings of the 10th CCRTS: The Future of Command and Control*, Decisionmaking and Cognitive Analysis, 1-30.

[26] Kahneman, D., and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology* 3:430-454.

[27] Convertino, G., Billman, D., Pirolli, P., Massar, J.P., and Shrager, J. (2008). The CACHE study: Group effects in computer-supported collaborative analysis. *Computer Supported Cooperative Work* 17:353-393.

[28] Lehner, P.E., Adelman, L., Cheikes, B.A., and Brown, M.J. (2008). Confirmation bias in complex analyses. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Human* 38:584-592.

[29] Kretz, D.R., Simpson, B.J., and Graham, C.J. (2012). A game-based experimental protocol for identifying and overcoming judgment biases in forensic decision analysis. In: *Proceedings of the 2012 IEEE Conference on Technologies for Homeland Security*, pp. 439-444. Waltham, MA: IEEE.

[30] Nickerson, R.S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175-220.

[31] Wheaton, K. (2014). Reduce bias in analysis: Why should we care? Retrieved from http://sourcesandmethods.blogspot.com/2014/03/reduce-bias-in-analysis-why-should-we.html.

[32] Rottenstreich, Y., and Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review,* 104(2):406-415.

[33] Tversky, A., and Koehler, D.J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review,* 101(4):547-567.

[34] Ayton, P. (1997). How to be incoherent and seductive: Bookmakers' odds and support theory. *Organizational Behavior and Human Decision Processes,* 72(1):99-115.

[35] Fox, C.R., Rogers, B., and Tversky, A. (1996). Option traders exhibit subadditive decision weights. *Journal of Risk and Uncertainty,* 13(1):5-19.

[36] Mandel, D.R. (2005). Are risk assessments of a terrorist attack coherent? *Journal of Experimental Psychology: Applied*, 11(4):277-288.

[37] Dhami, M.K., and Mandel, D.R. (2013). How do defendants choose their trial court? Evidence for a heuristic processing account. *Judgment and Decision Making,* 8(5):552-560.

[38] Wallsten, T.S., Budescu, D.V., and Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science,* 39(2):176-190.

[39] Baratgin, J., and Noveck, I. (2000). Not only base-rates are neglected in the Lawyer-Engineer problem: An investigation of reasoners' underutilization of complementarity. *Memory & Cognition,* 28(1):79-91.

[40] Macchi, L., Osherson, D., and Krantz, D.H. (1999). A note on superadditive probability judgment. *Psychological Review,* 106(1):210-214.

[41] Mandel, D.R. (2008). Violations of coherence in subjective probability: A representational and assessment processes account. *Cognition*, 106(1):130-156.

[42] Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., and Fox, C.R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(3):573-582.

[43] Mandel, D.R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Frontiers in Psychology,*6(387):1-12.

[44] Osherson, D., and Vardi, M.Y. (2006). Aggregating disparate estimates of chance. *Games and Economic Behavior*, 56(1):148-173.

[45] Predd, J.B., Osherson, D.N., Kulkarni, S.R., and Poor, H.V. (2008). Aggregating probabilistic forecasts from incoherent and abstaining experts. *Decision Analysis*, 5(4):177-189.

[46] Tsai, J., and Kirlik, A. (2012). Coherence and correspondence competence: Implications for elicitation and aggregation of probabilistic forecasts of world events. In: *Proceedings of Human Factors and Ergonomics Society 56th Annual Meeting*, 313-317. Thousand Oaks, CA: Sage.

[47] Wang, G., Kulkarni, S.R., Poor, H.V., and Osherson, D.N. (2011). Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis*, 8(2):128-144.

[48] Clemen, R.T., and Winkler, R.L. (1999). Combining probability distributions from experts in risk analysis. Risk Analysis, 19(2):197-203.

[49] Baron, J. (1985). *Rationality and Intelligence*. Cambridge, UK: Cambridge University Press.

[50] Nelson, J.D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact and information gain. *Psychological Review*, 112(4):979-999.

[51] Mellers, B.A., Baker, J.D., Chen, E., Mandel, D.R., and Tetlock, P.E. (2017). How generalizable is good judgment? A multi-task, multi-benchmark study. *Judgment and Decision Makin,* 12(4):369-381.

[52] Bruine de Bruin, W., Fischbeck, P.S., Stiber, N.A., and Fischhoff, B. (2002). What number is "fifty-fifty"?: Redistributing excessive 50% responses in elicited probabilities. *Risk Analysis*, 22(4):713-723.

[53] Armstrong, J.S. (2001). Evaluating forecasting methods. In: *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Armstrong, J.S. (Ed.). Norwell, MA: Kluwer.

[54] Willmott, C.J., and Matsuura, K. (2005). Advantages of the mean absolute error (*MAE*) over the root mean square error (RMSE) in assessing average model performance. *Climate Research,* 30(1):79-82.

[55] Cohen, J. (1992). A power primer. *Psychological Bulletin,* 112(1):155-159.

[56] Brenner, L.A., and Rottenstreich, Y. (1999). Focus, repacking and the judgment of grouped hypotheses. *Journal of Behavioral Decision Making,* 12(2):141-148.

[57] Tetlock, P.E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.

[58] Grissom, R.J., and Kim, J.J. (2005). *Effect Sizes for Research: A Broad Practical Approach*. Mahwah, NJ: Erlbaum.

[59] Ruscio, J., and Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research,* 47(2):201-223.

[60] Vargha, A., and Delaney, H.D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics,* 25(2):101-132.

[61] Tetlock, P.E., and Lebow, R.N. (2001). Poking holes in counterfactual covering laws: Cognitive styles and historical reasoning. *American Political Science Review,* 95(4):829-843.

[62] Friedman, J.A., Baker, J.D., Mellers, B.A., Tetlock, P.E., and Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly*, 62(2),410-422.

[63] Mandel, D.R., and Barnes, A. (2018). Geopolitical forecasting skill in strategic intelligence. *Journal of Behavioral Decision Making,* 31(1):127-137.

[64] Prims, J.P., and Moore, D.A. (2017). Overconfidence over the lifespan. *Judgment and Decision Making,* 12(1):29-41.

[65] Ashton, A.H., and Ashton, R.H. (1985). Aggregating subjective forecasts: Some empirical results. *Management Science,* 31(12):1499-1508.

[66] Libby, R., and Blashfield, R.K. (1978). Performance of a composite as a function of the number of judges. *Organizational Behavior and Human Performance,* 21(2):121-129.

[67] Winkler, R.L., and Clemen, R.T. (2004). Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis,* 1(3):167-176.

[68] Surowiecki, J. (2004). *The Wisdom of Crowds*. New York, NY: Doubleday.

[69] Mandel, D.R. (2015a). Accuracy of intelligence forecasts from the intelligence consumer's perspective. *Policy Insights from the Behavioral and Brain Sciences* 2:111-120.

[70] Stastny, B.J., and Lehner, P.E. (2018). Comparative evaluation of the forecast accuracy of analysis reports and a prediction market. *Judgment and Decision Making,* 13(2):202-211.



Chapter 13 – INTELLIGENCE AND THE ANALYSIS OF NARRATIVES

Peter de Werd[1]
Netherlands Defence Academy
NETHERLANDS

13.1 INTRODUCTION

The academic International Relations (IR) subfield concerned with the study of intelligence, or 'the second oldest profession' with its expert practice of intuitive operational and analytic tradecraft, is relatively young [2]. It was only after the Second World War that Western intelligence became more fully institutionalized in agencies that served as comprehensive 'libraries for national security' [3], p.5. Parallel to the bureaucratization of intelligence, technical developments such as the invention of radar or programmable crypto decoding machines had a significant impact on the field. The mathematicians and other academics that were hired to work with (signals) intelligence contributed to an academic professionalization of intelligence.[2] Parallel to various social sciences, the positivist scientific ideal of the natural sciences was pursued in intelligence. This was reflected, for example, in J. David Singer's iconic mathematical threat assessment model (*threat = estimated intentions x estimated capabilities x observed activities*) [9]; [10]. The ultimate aim was to find the objective ground truth. Until today this positivist empiricist paradigm has remained dominant in the practice and study of intelligence [11], pp. 33-34; [12]; [13]. Building on different, critical philosophical ground, this chapter seeks to provide a rough outline of a discursive approach to intelligence that holds potential for the practice of intelligence analysis. Rather than contesting positivism, the idea is to move beyond the limitations of the dominant paradigm in intelligence and pursue a more holistic approach that also recognizes interpretivist aspects of social reality. First, this chapter introduces the philosophical theory of critical realism and the related view on causation. Second, this chapter outlines corresponding theoretical components of critical discourse analysis and securitization theory that combine into a methodology of *Analysis by Contrasting Narratives.* Lastly, besides advocating the added value of such an approach, the chapter addresses challenges and anticipates potential critique from intelligence professionals.

13.2 CRITICAL REALISM, CAUSALITY, AND THE ROLE OF LANGUAGE IN SECURITY AND INTELLIGENCE

Critical realism, a contemporary and critical form of realism rooted in the work of British philosopher Roy Bashkar, serves as a theoretical 'middle ground' that transcends the causation-constitutive or explaining-understanding divide, and the structure versus agency debate [14]; [15]; [16]; [17]; [11], p. 49; [18]; [19]. For decades, these distinctions have entrenched positivist empiricists and post-structuralist reflectivists, particularly in IR. For positivists, causal relations have been limited to efficient ('pushing and pulling') regularity relations of observables, as described in Ref. [16], location 169. Research has involved pattern seeking as an additive approach to 'stack' isolated singular causes (causal mechanisms) in 'closed systems'. This has placed logical determinism (or discovering laws) as the central aspect of the scientific endeavor. In this respect, the works of British philosopher David Hume on causation have been highly influential in shaping the core principles of positivist empiricist positions on causality [20]. Post-structural reflectivist approaches that critique Humeanism have focused on understanding how ideational aspects (ideas, norms, conventions, and discourses) are constitutive of the social world. However, in the rejection of

[1] This chapter reflects a segment of Peter de Werd's "Critical Intelligence Studies? A Contribution" [1].

[2] For example, see Refs. [4], [5], and [6], also described in Refs. [7], pp. 25-28, and [8].

Humean causality and avoidance of the terminology, these approaches are also unnecessarily 'reductionist' as they exclude materialistic and deterministic analysis.

Critical realism acknowledges a form of interpretivism and thus moves explicitly away from the objectivist/empiricist paradigm. In contrast to post-structuralism, however, critical realism views observation and interpretation as a matter of epistemology, not ontology. In other words, there is a 'real' world out there. Reality is regarded as differentiated.

> *[C]ritical realists distinguish the real from the actual and the empirical. The 'real' refers to objects, their structures or natures and their causal powers and liabilities. The 'actual' refers to what happens when these powers and liabilities are activated and produce change. The 'empirical' is the subset of the real and the actual that is experienced by actors. Although changes at the level of the actual (e.g., political debates) may change the nature of objects (e.g., political institutions), the latter are not reducible to the former, any more than a car can be reduced to its movements. Moreover, while empirical experience can influence behaviour and hence what happens, much of the social and physical worlds can exist regardless of whether researchers, and in some cases other actors, are observing or experiencing them.* [21], p. 204

Critical realists do not conceive of the world in terms of either-or. Instead of viewing structure and agency as antithetical, critical realists hold that they conflate in a dialectical relationship. Reality is an 'open system' in which causal powers interact, enforce, or counter each other. Whether causal powers in the domain of 'real' become active at the 'actual' level and can be observed as a fact in the 'empirical' domain depends not only on the social conditions that enable the activation of causal forces, but also whether other powers work against it [22]. Causal powers can be active, dormant, or countered.

Rather than avoiding the term 'causality', it is possible from a critical realist philosophical position[3] to rethink and reconceptualize causality beyond the traditional positivist empiricist Humean account of observable constant conjunctions [16]. Empiricism is not the only way of gaining knowledge. Unobservable processes of social construction can be understood by interpreting motives, reasons and meanings, ideas, rules, norms and discourses, and the way these are influenced by the social context. Causes interact and reflect other causes. Instead of additive analysis of singular causal mechanisms, the complexity of the social world requires an integrative approach. It is necessary to consider the 'network of causality' or 'causal complex', rather than single out an individual causal mechanism [23], pp. 18, 22-23, 47; [24]; [25]; [26], p. 21; [16], location 1935. Causes cannot be considered mechanisms, although causal processes or interactions of causes could perhaps arguably be considered in such a way. Critical realists do not necessarily reject the term 'mechanism', but in this chapter it is best avoided to reduce confusion over Humean associations.

The question is how to trace and analyze social conditions and powers to explain social processes in a causally adequate manner. Following German IR professor Alexander Wendt and others, Aberystwyth University International Politics scholar Milja Kurki has made a fruitful effort to explore the use of Aristotle's four-fold conceptualization of causes from a critical realist perspective [16]; [26], pp. 1-22; [27]; [28]. Without attempting to address all of Aristotle's theorizing, Kurki demonstrates that the typology of material, formal, efficient and final cause is instrumental to specifying the concept of 'causal complex' and identifying how multiple types of causes interact.

First, *material cause* relates to the nature and properties of matter that enable and constrain possibilities of social action (in what way and for what matter can be used). Material cause is more than substance, as it also encompasses artefacts. At a secondary level, matter can hence be thought of as formed objects with a passive potentiality that shapes basic conditions of social reality [16], location 2604. Without (materials to make) weaponry and bombs, there is no capability to act with violence. However, the causal power of matter is also

[3] Ontological philosophical realism and epistemological interpretivism.

intertwined with the physical and conceptual arrangement or social structure in which it is used. Weapons possessed by a friendly entity hold different meaning, or potential, than those owned by an adversary. *Formal causes* relate to the relatively stable ideational context that generates functional shapes of appearance. Ideas, conventions, norms and discourses affect the ways in which meanings are defined, articulated, circulated and conceived. Material causes can influence formal causes, as property can increase social status in some contexts. Conversely, a national security discourse can result in the (defensive) organization of infrastructure. Both types of causes can be thought of as constitutive conditions or structures that enable and constrain possibilities for action or agency. They form 'related wholes within which intentional actors act and thereby reproduce or transform the facilitating social conditions (material and formal) of their own activity' [16], location 2693.

Efficient causes are what is generally conceived as causality. It is the entity or actor that activates movement, interaction, and change. It brings about actions that reflect, recreate, and transform matter and form. This does not only relate to physical action: while discourses can be viewed as constitutive in terms of formal causes, discursive action also has efficient causative effects. By making (provocative) statements in certain settings, specific articulated meanings can become actualized and influence social reality. *Final causes* are teleological. What are the motivations, visions, intentions, or reasons for action? The purpose of action is related to efficient causes, but distinct. Of course, the intended effect of actions can differ from what they actually cause. In case of discourses, the underlying motivation or purpose recipients read into statements may differ widely from what a producer of texts has intended.

These four categories or types represent both constitutive structures (material and formal cause) and causative agency (efficient and final causes), or facilitating conditions and drivers. The various types of causes interactively generate, counter, enable, or constrain effects in the social world. Actions are related to a purpose, but also situated in an ideational and material context. Any scientific methodology that relates to critical realism needs to be able to account for and reflect upon the activation of potential powers (or 'potentialities') against the backdrop of the distinction between deeper intransitive social structures and more transitive social practices and events. They have to acknowledge causal pluralism and approach a *complex of causes* holistically to study social phenomena.

Ontologically, the real world 'out there' consists of objects with 'real properties and causal powers by virtue of their composition' [16], location 2316. Scientists can make plausible causative statements as they study the nature and role of a plurality of causal powers that create social reality.

> *Causes, for philosophical realists, are not equated with regularities but can be seen to refer to real ontological features of the world. Scientific causal explanation, then, is not equated with analysis of observable regularities, but is seen to arise from the construction of conceptual models that try to grasp the nature of objects through making existential claims about their constituting structures and causal powers, thereby enabling explanations of various 'actual' or empirical processes and tendencies. Regularities are of interest to science because they allow us to test theories regarding causal powers in artificial closed system environments. Yet, observed regularities do not constitute causality: causality exists in the underlying causal powers and causal explanation in accounting for these underlying causal powers.* [16], location 2322-2326

Because not all can be observed, observable regularities are neither necessary nor adequate to explain causal relations. This has implications for the way knowledge is gained. As Peter Gill and Mark Phythian describe, the creative process of abduction or redescription offers a way to find new connections by adopting and testing hypotheses about socially produced realities. Abductive research entails the process of accepting causality between certain social structures, processes, and events, and in addition reflecting upon this relation as the research progresses. By assuming the social conditioning of a society by dictatorship and globalization, political debate through demonstrations and other utterances generates meaning about the essence and workings of this conditioning. However, although the theoretical framework to study these

situated demonstrations and utterances in context provides insights that are valid knowledge, it does not encompass *all* there is to know about the causal forces at play.

> *Neither deduction nor induction alone is adequate in social science: we do not 'discover' new events but we do discover new connections and relations that are not directly observable and by which we can analyze already known occurrences in a novel way. […] By applying alternative theories and models in order to discern connections that were not evident, intelligence scholars […] [are not] merely describing reality as if through a clear pane of glass: they are seeking to make sense and thus actively 'create' the worlds of intelligence, government and IR* [11], p.40.

Ultimately, our knowledge is imperfect as it is 'theory-laden' [29]. Hence, another implication of the critical realist philosophical position is the importance of a reflexive scientific attitude. Scientists are required to continuously consider the extent to which research designs have been constrained or influenced by social and political conditions. The same is applicable for the practice of intelligence. The role of the intelligence analysts and intelligence consumers (the 'observers') in society must be made explicit when analyzing intelligence problems. Intelligence analysis is not 'value-free', but socio-politically situated [11], p.39. Reflection on how situated intelligence shapes problems in particular ways thus turns into an integral part of intelligence analysis. This reflexive attitude also enables deeper analysis of intelligence problems. Problematizing the situatedness of intelligence analysis itself opens up possibilities of a deeper understanding of what contexts, settings, and perceptions of security drive the actions and shape the motivations of intelligence subjects.

A sense of the added value of a critical realist scientific approach to the predominantly positivist practice of intelligence analysis is gained by examining the 'Intention, Capability, Activity' (ICA) framework [10]. The multiplicative 'mathematical' threat assessment model is highly appreciated in intelligence analysis, for example when analyzing terrorist threats. When comparing the Aristotelian typology of causes to this framework (intention – final cause, capability – material cause, activity – efficient cause), the formal cause is evidently backgrounded as a distinct analytical category. One could argue the category 'intention' indirectly refers to motivation and worldview, and hence includes the cultural or ideational context; but the point is that formal causes are fundamental facilitating conditions that color the meaning of actions. In addition to the question 'are there terrorists out there that target us?', the question 'why?' deserves more attention. What actors and audiences are involved, at what level and how are they socio-politically (and culturally, religiously, etc.) situated?

13.2.1 Critical Discourse Analysis

As a fundamental part of the causal complex, the study of discourse structures within socio-political ideational contexts encompasses a relevant approach to intelligence and security. The use of language is a primary semiotic mode of entry to study objects of research and the social dynamics of causal complexes. Not all is discourse, however. Non-semiotic (or 'extra-discursive') elements also make up social reality, such as the material world, people, social relations and action. Critical realism underlines the significance of both semiosis and non-semiotic elements [21], p. 219. This is also a theoretical stand explicitly reflected in Critical Discourse Analysis (CDA) as opposed to other forms of discourse theory (see, for example, Refs. [30] and [31]). CDA as introduced by British linguistics professor Norman Fairclough is a dialectical-relational discursive approach that recognizes how (social) structure and agency (both discursive and non-discursive) influence each other. It is impossible to see discourse as a separate object outside of context. The challenge in discourse analysis is to recognize all types of causes and be open to working in an interdisciplinary or transdisciplinary way, while preventing the tendency among social scientists to overdetermine non-textual aspects [32], pp. 294-295. In the Aristotelian idea of the causal complex, materiality is primary. However, this is not antithetical to Fairclough's approach. Rather, in indicating the significance of both material and discursive aspects of causal complexes, they balance each other.

Compared to other variations of critical discourse analysis, Fairclough's approach excels in his sociological approach (see Refs. [31], pp. 101-128; [33]; [34]). It connects different levels of analysis: social structures (e.g., nations, the global network society), social practices (e.g., international politics, the information society) and social events (statements, actions, occurrences). These correspond to a discursive division between language, orders of discourse (e.g., national security) and the discursive practice of production and consumption of texts (e.g., a declaration of war) [32], [35], [36]. Texts not only include written or spoken language but also symbols, signs, images, music or any other means to communicate. *Narratives*, or discourses, combine basic elements in texts (events, actors, timeframes, locations) to form stories that advance a certain meaning and logic [37], [38], [39]. They reflect and (re)create their social conditions of production and interpretation, in terms of texts, interactions and context [40], pp. 57-58. Identifying the context of text production and consumption particularly allows identifying and situating narratives in distinct social domains. This forms the basis for the methodology of Analysis by Contrasting Narratives.

CDA is critical as it is concerned with 'contradictions between what is claimed and expected to be and what actually is' [40], p.9. Discourse is explained with regard to how such contradictions, tensions, and articulations of social difference are (necessary) elements of a broader social reality of which they are part. How does power make or keep certain meanings or ideologies dominant, and how do dominant meanings or ideologies maintain power relations and difference? Apart from analyzing discourse, CDA includes reflecting on the interpretations, evaluations, critiques, and explanations of discourse participants. How are they subject to relations of power? Through self-consciousness and self-reflexivity, the researcher is separated from discourse participants when interpreting discourse. By systematically analyzing all types of causes, the explanations supersede the social reality of the discourse or narrative. This is different from normative critique of discourse displayed in narratives or counter-narratives within the same or a different social domain [40], p.12.

Discursive analysis thus encompasses the study of production and consumption of multiple texts within contextualized discursive practices that are woven together to form discourses or narratives. Texts are constituent and constitutive, they reflect and create reality. But they do so while production and consumption are subject to relations of power. Who has the means and authority to say what about whom or what, or to keep others from speaking in this respect? Who has the ability and willingness to listen? Non-discursive power relations enable and constrain entities to produce, combine, reproduce or consume texts. CDA studies the forming and workings of knowledge and power through language against the background of social reality. It enables identification of distinct narratives in different social domains. What should be looked for in narratives? How and for what are power relations enacted or activated? The concept of securitization provides the necessary focus to analyze narratives.

13.2.2 Securitization

The debate in critical security studies on securitization theory is highly significant to intelligence analysis. The theorizing on this critical concept of security is part of an effort to broaden and deepen the scope of traditional concepts in security studies [41]. Security is not defined as an objective reality, but as a social construct or a form of interpretation manifested in the use of language. The traditional positivist paradigm limits security to the survival politics of states, primarily with a military focus. In general, the concept of securitization is associated with two generations of theorists. This chapter considers the first generation, represented by the Copenhagen School, too limitative [42]. They narrowly define securitization as a speech act, centralizing the utterance itself. As a felicity condition, the audience has to accept the extraordinary status of a political issue for the securitization move to be successful. The main problem with this is that instances or processes of securitization are deemed 'discontinuous changes' or social 'quantum jumps' that have no preceding causal relations, for example to non-discursive aspects of reality [43]. How then, to trace and explain how it is possible securitization occurs and what social elements act as a causal force? The discursive speech act approach leaves enabling or constraining wider social conditions, underlying forces and non-discursive factors out of the analysis.

A second generation of securitization scholars argues securitization is more complex, dynamic, and nuanced (for example, see Refs. [23]; [44]; [45]. Some of the early first generation theorists have been deemed 'internalists' for their emphasis on the speech act itself, whereas second generation theorists were called 'externalists' [46]. Another distinction often made is between a philosophical and a sociological approach. Belgian International Relations Professor Thierry Balzacq is a leading figure in the second generation. His concept of contextual sociological securitization is grounded in critical realism [23]; [44]; [47]. He deems securitization not a speech act, but a pragmatic act. Securitization is:

> *an articulated assemblage of practices whereby heuristic artefacts (metaphors, policy tools, image repertoires, analogies, stereotypes, emotions, etc.) are contextually mobilized by a securitizing actor, who works to prompt an audience to build a coherent network of implications (feelings, sensations, thoughts, and intuitions), about the critical vulnerability of a referent object, that concurs with the securitizing actor's reasons for choices and actions, by investing the referent subject with such an aura of unprecedented threatening complexion that a customized policy must be undertaken immediately to block its development.* [23], p. 3

This definition implies the use of language is explained within certain contexts, rather than as utterances of a sovereign speaker to a sovereign hearer. Also, securitization can exist in practices other than words, such as bureaucratic procedures or technologies – an approach that clearly conforms more to this chapter's philosophical underpinnings on causal relations. Balzacq draws on Dutch language scholar Jacob Mey's theory of pragmatic acts:

> *The theory of pragmatic acts […] does not try to explain language use from the inside out, that is, from words having their origin in a sovereign speaker and going out to an equally sovereign hearer […]. Rather, its explanatory movement is from the outside in: the focus is on the environment in which both speaker and hearer find their affordances, such that the entire situation is brought to bear on what can be said in the situation.* [48]

Balzacq's main critique of the Copenhagen School is built on three assumptions [47]. First and foremost, he assumes that securitization as speech act leaves the status of the receptive, predefined audience unaccounted for. He recognizes the practical difficulties with identifying and analyzing audiences, but nevertheless argues the concept of audience requires differentiation [49]. Hence, Balzacq proposes a distinction between formal and moral support of securitization [47].[4] Formal support comes from the audience that provides the necessary legitimate mandate to execute special measures to deal with the threat. Moral support conditions formal support and securitizing actors strive to prompt a moral audience as large as possible to strengthen social relations and their position of authority. For this the securitizing actor has to take into account the audience's frames of reference, their readiness to be convinced (depending on their trust in the securitizing actor) and the ability to (indirectly) grant or deny a formal mandate [47]. Others have a wider view and consider 'various types and parallel' audiences that relate to different (general or specific) functions of securitization processes, such as raising an issue on the agenda, reproduction of a certain security status or legitimizing past and future actions [51]; [52].

The position supported in this chapter is that identifying audience assent and declaring securitization 'successful' is not necessary to analyze the causal relations and effects involved. Audiences are an essential component of securitization, but not necessary in terms of causal determinacy, granting deontological powers to the securitizing actor. Securitization efforts reflect the causal *adequacy* of audiences. Empirical evidence on audiences and their responses might be fragmentary or absent, but that does not *a priori* imply valid conclusions cannot be inferred on audiences; for example, from alignment of securitization efforts with particular social structures and practices. Potentially, research on complex intelligence problems can thus also contribute to the debate on securitization within critical security studies in its analysis of various types of audiences associated with the narratives or resonating with the securitization efforts. Most securitization research concentrates on

[4] A point developed further by Paul Roe in Ref. [50].

(democratic) institutionalized environments in which power relations have been established to a certain extent [53]; [54]. Based on case studies of narratives related to complex intelligence problems, the role of various types of audiences in different social domains need to be analyzed more widely. This includes less institutionalized environments, such as social networks. Not to filter out singular essential characteristics but to generate more insights on the nature and status of securitization audiences in these different contexts.

A second aspect of Balzacq's critique concentrates on how the use of security modifies the context, yet in order to be 'effective' such use must be aligned with an external context that is independent from the use of language. In other words, there is a distinction between the situational context of the securitization effort and the background context or *zeitgeist*. Different audiences (both moral and formal) find themselves situated in various settings (for example, popular, elite, technocratic, scientific, religious) [55]; [23], p. 37; [56]; [57]; [58]. Each of these settings can be characterized by particular expectations, specialized language, conventions, and procedures. Hence, while some audiences in some settings might resonate with a securitizing actor and his efforts, other audiences in other settings might not. The situational context can also be characterized by circumstances, such as a large-scale natural disaster, that make other securitized threats relatively less important. The concept of settings emphasizes that what entails 'security' or a threat to it differs over time and space, and the practical effects of securitization efforts depend on the situatedness of the security discourse itself [59]. But ultimately, these more dynamic settings can be situated in more durable social structures, such as institutions, class, and culture. For example, the extent to which a society is sensitive to xenophobia can relate to the degree of geographical, social or technological isolation. Discourses or narratives that encompass securitization efforts do so with respect to different audiences within an overarching social structure.

Third, the power of securitization efforts is related to the social position of the speaker and his unequal access and ability to use discursive resources. This is part of what Michel Foucault refers to as a wider 'system of relations' or 'dispositif' that links 'a thoroughly heterogeneous ensemble of discourses, institutions, architectural forms, regulatory decisions, laws administrative measures, scientific statements, philosophical moral and philanthropic propositions – in short, the said as much as the unsaid' [60]. Securitization is subject to constellations of power, but Balzacq holds it is also the language itself that has 'an intrinsic force that rests with the audience's scrutiny of truth claims, with regard to a threat, made by the speaker' [47], p. 173. A securitizing actor needs to use the appropriate words that fit the frames of reference of audiences to win support. In sum, Balzacq's pragmatic model of security specifically aims to accommodate analysis of 'the psycho-cultural orientation of audiences, the wider context, and the differential power between the speaker and the listeners' as key aspects [47], p. 173.

Summarizing his critique on practically excluding the context, status, and nature of audience, Balzacq defines three dimensions or levels of analysis that provide different perspectives on securitization: agents, acts and context [23], pp. 35-37. The level *agents* encompasses the various actors, including audiences, the personal and social identities and the power relations involved. The discursive and non-discursive practices that endorse securitization are the focus of the second level: *acts.* This includes the type of language used, the strategic use of heuristic artefacts as social devices to generate the conditions that enable the mobilization of audiences, the *dispositif* of (or generated by) the securitization process and the customized policies generated by securitization. The third level, *contexts*, refers to the way securitization is situated socially and historically in situational and wider background contexts. Among second generation securitization theorists, academic debate is also ongoing about the nature of reverse processes of securitization. There is always a potential for any debate to 'open up', unmake, desecuritize, or transcend [44]. Discussions on the nature and workings of several logics or strategies of desecuritization and contesting security, such as resistance, emancipation, and societal or organizational/infrastructural resilience, are relevant for developing the ACN methodology, but fall outside the scope of this chapter to discuss the philosophical theoretical foundation and situate the approach in intelligence studies [44]; [61], pp. 167-185.

Balzacq's securitization theory and Fairclough's theoretical discourse model complement each other and enable one to *identify* and *analyze* distinct narratives. The differentiated critical realist view of reality and

Aristotle's four-fold typology of causes provide for the broader outline or philosophical theoretical foundation, grounding critical discourse analysis and securitization theory. The latter theoretical components enable one to zoom in on the dialectical relation between social events, practices, and structures. In and through discursive and non-discursive action, the manifestation and workings of power relations and processes of identification take shape. Consistently attributed meanings in service of maintaining power relations and articulating difference between a particular 'self' and others can gradually transform into ideology, or even become generally accepted 'common sense'. This process of naturalization describes how the various types of causes as defined by Aristotle relate, interact and also transform each other [40], pp. 35, 126; [36], p. 218; [35], pp. 9, 27. CDA and securitization provide the theoretical components and logic to trace how various facilitating conditions and drivers (or causes) combine, and explain how such causal complexes affect social realities.

13.3 ANALYSIS BY CONTRASTING NARRATIVES

The following rough outline of the Analysis by Contrasting Narratives (ACN) methodology involves a practice that to some intelligence practitioners might seem quite radical. However, as the intelligence environment has become increasingly complex, new ways of thinking are required more than the rearranging responsibilities and structures often proposed after intelligence failures. Rather than problematizing any 'duplication of function' between intelligence analysis and policy analysis, a *more* integrative (or hybrid) approach is called for [62]. This partly is a rendition of the critical self-reflection on the socio-political situatedness of intelligence analysis and recognizes how the perspective and actions of the intelligence consumer essentially shape the intelligence problem. ACN advances cooperation and jointness among intelligence professionals and working-level policymakers as they are necessary to contribute to holistic sensemaking of complex intelligence problems. The strategic narrative of the intelligence consumer is incorporated in the overall analysis. It is possible for trusted outside experts to become involved in analyzing additional narratives if the necessary expertise is lacking in intelligence organizations to function as situated (knowledgeable) critical (self-reflexive) interpreters.

Such approaches are also relatively uncommon in security studies.[5] Research on securitization has mostly focused on the use of language in *a specific* discourse. The researcher takes a normative stance and deconstructs relations of power and meaning for this discourse. However, social events become part of *various* discourses or narratives as they are interpreted in various ways by different entities and at different levels. What makes the ACN approach distinct is that *several* discourses (at least three) are compared and contrasted, instead of studying and deconstructing one 'dominant' narrative. It seeks to identify *basic analytic narratives* for various entities that manifest at different levels and dominate the attribution of meaning, especially in terms of securitization. *Who* has *what power*, on *which level*, in *what setting*, over *what audience*, to attribute *what meaning of security*, in *what texts*, which support *what interests* and motivates *what action*, *by whom*? These multiple narratives relate to several of a set of social events (such as a terrorist attack, a declaration, or troop movements) that are part of an intelligence problem.

The terms *basic* and *analytic* are used to underline how defining the level and contours of any discourse is ultimately a matter of choice for the researcher that needs to be made explicit [64]. For example, it is possible to define a United States institutional discourse on *Al Qaeda* at the national level as the relevant strategic narrative of the intelligence consumer. But at times, defending its cohesion can prove to become difficult as different readings of events at the subnational level amount. The attack on the US Embassy in Benghazi in 2012 led Republicans like House Intelligence chair Mike Rogers to conclude it was a 'coordinated, military style commando-type raid', a 'pre-planned, organized terrorist event' that was 'led by *Al Qaeda*' [65]. The Democratic US Ambassador to the UN Susan Rice stated shortly after the attack it was actually a street protest against an anti-Muslim video on YouTube that got out of hand, and some extremist

[5] One of the exceptions that considers two perspectives is Stritzel and Chang's "Securitization and Counter-Securitization in Afghanistan" [63].

elements had eventually joined [66]; [67]; [68]. Later, *New York Times* journalist David Kirkpatrick added another reading of the events [69].

Deciding on the relevance of narratives and determining the adequate level of analysis entail a process of redescription, or abductive reasoning. As social orders and narratives are defined, new insights from ongoing analysis might lead to redefine or adjust the level or scope of analysis. To ensure a widening of understanding, it is important to ensure two or more narratives relate to essentially distinct social practices (e.g., international politics, investigative journalism in the global network society or Salafi-jihadism). Each social practice enables distinctly different (non-)discursive practices such as the articulation of foreign policy, the publishing of memoires, freelance research, or legitimization of violent actions in Arabic genres such as memoranda declarations (*hukm*) or decrees (*fatwa*). In each narrative bombings and troop deployments are reflected on in a specific way, (re)creating different processes of securitization and identification in co-existing realities. Some aspects of these narratives are secret or concealed, others are public, but in all cases, securitization requires an audience (albeit one that is sometimes hidden or compartmentalized).

Securitization efforts are about the activation of causal complexes to generate social effects. Narratives of antagonists or other entities, as well as the strategic narrative of the intelligence consumer might reflect such efforts. These are termed 'macro narratives'; in and through which the central entities have considerable discursive power over audiences, and are able to engage in extraordinary security practices, hence impacting a complex intelligence problem. In addition, the ACN methodology also involves incorporating one or more 'micro narratives'; accounts that often reflect critically on securitization efforts in the macro narratives, while their producing entities lack power to act in terms of security. The latter are narratives that can be situated in yet another social domain. Because micro narratives exist relatively outside of the dominant social orders of the macro narratives, they have the potential to function as 'commentators' and highlight tensions, inconsistencies or contradictions in macro narratives. Such insights are valuable to trace multi-consequentiality of statements and actions across social domains: did securitization efforts in one narrative have an effect in other social domains? Multi-consequentiality of securitization efforts can be illustrated as follows (Figure 13-1).

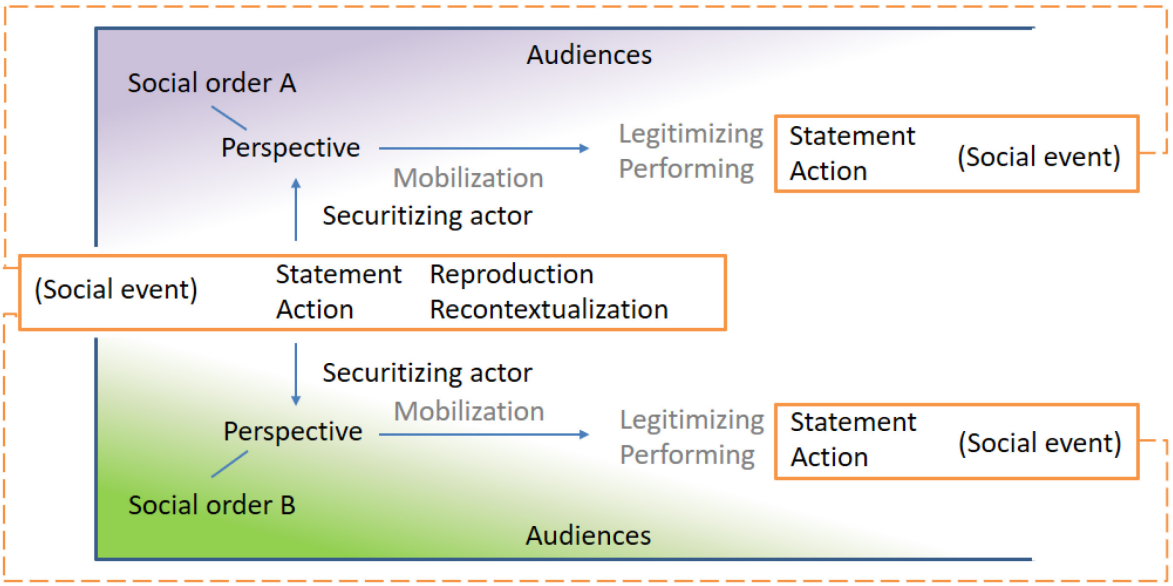


Figure 13-1: Aspects of Social Events Serve, Within a Certain Social Context, as Heuristic Artefacts to Mobilize Particular Perspectives That Enable New Social Events to Occur. Any (new) event is potentially multi-consequential.

ACN does not imply that traditional methods and analytic concepts are obsolete. However, traditional 'puzzle solving' becomes more relative as ACN allows analysts to highlight multiple perspectives at an

initial stage and escape the limitations of the political strategic narrative that intelligence consumers use to legitimize their own policy [70]. It is more important to seek various meanings of discourses than to increase collection of rhetoric. Many intelligence analysts believe that the more means an intelligence organization has the closer it gets to 'the truth'. This is debatable. The effort by the US to achieve 'total information awareness' by primarily focusing on increasing the quantity of available (meta)data seems of only limited use, as from an ACN perspective the 'causality' of correlations found in big data with poor content and context can be questioned. The methodology of ACN points the way in a different direction. It is about understanding various meanings of events and circumstances, by situating statements and actions in various causal complexes and explaining their effects on social phenomena (see Figure 13-2 as an example).

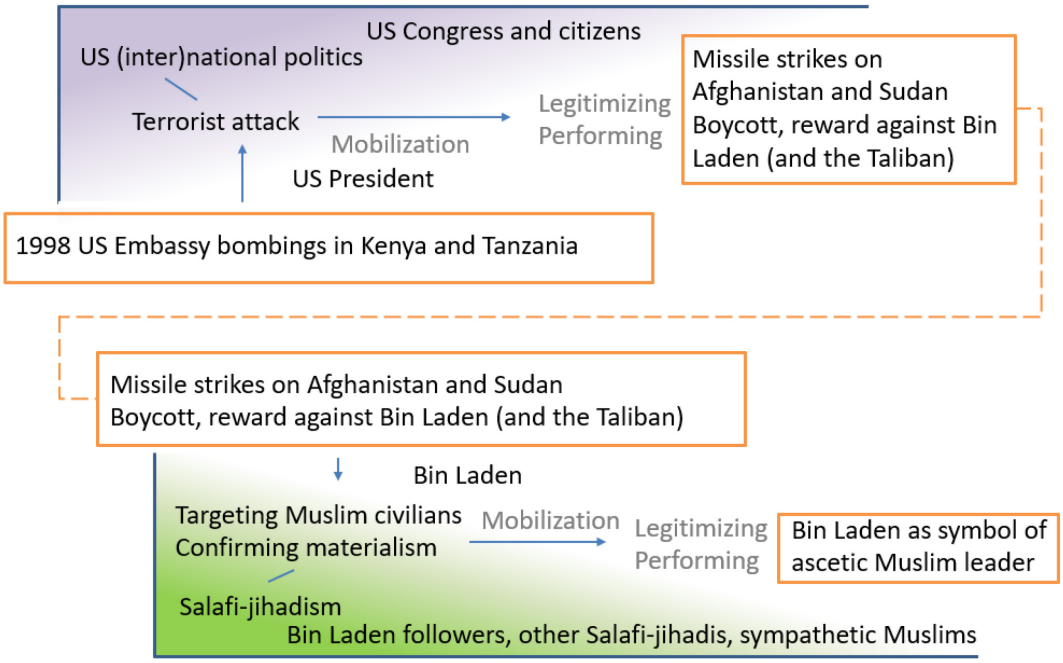


Figure 13-2: An Example of ACN. Whereas US President Clinton's response to the 1998 terrorist attacks in Africa represented a holistic approach for his US audiences, for Salafi-jihadis and other Muslims, it strengthened Bin Laden's narrative.

13.3.1 Possible Objections from the Field

Operationalizing securitization theory and critical discourse analysis for intelligence into ACN might receive criticism from two opposing fronts in intelligence studies. First, in the sphere of the study for intelligence, sceptics might argue that existing analytic methods are adequate to capture the multiple meanings entities attribute to (social) reality.[6] They would hold that various Structured Analytic Techniques (SATs) for intelligence analysis aim for reframing and improving accuracy of assessments by widening the analytical spectrum and assessing alternative points of view. In the (US) intelligence practice, SATs have a special status [71]; [72]. The set of techniques represents the ideal analytic tradecraft, or gold standard for analyzing intelligence problems. Yet to an important degree this status has been granted mostly by 'lore and assertion' [7], p.33. Intelligence scholars have recognized problems with assessing their actual rigor and efficiency [73]; [74]; [75]; [76]. Moreover, there are distinct differences between ACN and contrarian SATs, such as team A/B analysis or red hat analysis, devil's advocacy, premortem analysis and structured self-critique.

[6] A point raised by Stephen Marrin to the author.

At first glance, team A/B analysis seems suitable to grasp co-existing truths. Groups of analysts are formed that try to make the best case for different points of view or hypothesis with respect to the available data, providing policymakers with multiple assessments [71]. Yet, such analyses often lack the theoretical considerations that guide the *selection* of various points of view, and they have been known to enable policymakers to follow a preferred (or predefined) policy [77]. An infamous example is the intelligence failure of assuming that there were 'numerous' connections, 'areas of cooperation' and a 'shared interest and pursuit of weapons of mass destruction' between the Iraqi regime led by Saddam Hussein, and *Al Qaeda* and its leader Osama bin Laden [77], p. 155. This assessment was advanced by Policy Counter-Terrorism Evaluation Group (PCTEG), a team B group of analysts situated at the US Pentagon who deductively assumed there was a connection and sifted through raw intelligence and analytical products to find supporting evidence.

Thus, in contrast to ACN, team A/B analysis lacks the theoretical foundation informing the selection of perspectives, and the deeper distinction between various forms of causality. The team A/B practice also differs from ACN in that the various assessments are laid before the intelligence consumer to choose from. In ACN, the various perspectives are generated in a cooperative process that involves working-level policymakers to generate one of the narratives. Comparing and contrasting various narratives empowers analysts to ask additional questions and direct collection, while finding comprehensive insights that can be shared. Red hat analysis or red teaming is sometimes used as synonym for team A/B analysis [78], p. 243. However, it can also be viewed as a particularization, as it is performed by a group that is assigned to take on the perspective of the adversary when viewing a problem and projecting possible courses of action, asking how the adversary would respond to developments and actions. In this case, the critique on a lack of theoretical backing for the selection of the perspective seems less applicable. Still, an implicit danger lies in the division between 'red' and 'blue'. Red implies an opposition to the blue frame, similar to black versus white, and it is a characterization that is attributed from the 'blue' perspective. A particular red hat technique or 'tool' is called 'four ways of seeing' [72], p. 77. It involves contrasting how two entities see themselves and how they see each other. But apart from the remark that 'thorough research should be conducted' on these perspectives, the theory lacks on how to accomplish this [72], p.77. Moreover, although challenging the conventional wisdom is an important function, ACN differs in the way it accomplishes this. Rather than a narrative and an anti-narrative, ACN seeks to understand narratives associated by a multitude of actors. From a methodological viewpoint, identifying additional different macro and micro case studies can only be encouraged.

Devil's advocacy is more about reviewing process than content, while ACN is about generating a variety of content or perspectives as a basis for analysis. With the help of devil's advocacy, proposed analytic assessments are challenged by one or more analysts who have not been involved in the analysis. Performed at the discretion of the management of the intelligence organization, it primarily provides critical peer review of the analytic rigor of assessments. Relatedly, premortem analysis and structured self-critique are techniques that internalize the function of a devil's advocate in analytic teams [72], p. 223. The various SATs in themselves are valuable, and ACN does not aim to become a substitute for any of them. As such, it is more fruitful to discuss ACN in terms of its potential contribution to the SATs taxonomy that has been gradually developed within the US and other Western intelligence communities [71]; [72]; [78], pp. 19-25; [79].

At a different level, various scholars have also advanced 'critical thinking' in intelligence studies, which relates to simultaneously improving standards of reasoning (or *thinking about thinking*) and arriving at a conclusion (or *thinking*) [80]; [81]; [82]; [83]; [72]. In the self-reflexivity that is argued for, some parallels can be found to the theoretical critical approaches described in this article. An effort has been made to move beyond the practical day-to-day critical traits and techniques, to include a taxonomy of reasoning types. Combining critical traits and techniques with different types of reasoning (a taxonomy of abductive hypothesis testing, causal analysis, counterfactual scenario reasoning, and strategy assessment) would result in identifying the desired 'best explanations, relevant causes, probable scenarios, and optimal decisions' [84]. However, there are distinct differences compared to critical theory. Critical thinking is more of a short- or middle-range concept related to practical wisdom employed to improve intelligence analysis and education.

The concept relates to all aspects of the intelligence cycle (i.e., requirements, collection, analysis, and dissemination), but the limitations of this critical thinking become most visible with the aspects of intelligence requirements and dissemination.

Critical thinking initially leaves the reference frames *of the customer* out of the equation. It focuses on identifying and serving the right customers by answering their key questions and recognizing a broader context for analysis that supersedes *the analysts'* initial frames. The 'knowing your customer checklist' does point out that other interested parties the customer might go to might have different perspectives on the issue at hand [80]. Traditionally, debates on politicization of intelligence are cut short by adopting the view that both top-down 'cherry-picking' or bottom-up 'self-censorship' are 'unprofessional' [80], pp. 159-160. Objectivity and integrity are regarded as central to analysts, while it is the management of intelligence organizations that bears the responsibility to 'bring intelligence into the realm of politics without corrupting it' [3], p.77. At the stage of dissemination, the frame of the customer surfaces in the critical thinking literature as one of the criteria used to evaluate intelligence products. Is the product relevant to the customer's interests, and 'is it easy to translate it into the customer's frame of reference and responsibility' [80], p. 181?

Literature on the study for intelligence has much to offer, but the concept of critical thinking and the various SATs certainly leave room for ACN as an additional or alternative approach. ACN provides the theoretical foundation to ensure the distinctness of the various selected narratives or perspectives. It actively takes into account the socio-politically situated dominant strategic narrative that, because of its inherent abstraction and limitations, could steer analysis off course by creating blind spots. Securitization efforts, or critiques of them, provide a focus for these basic analytic narratives.

Furthermore, a critique of less immediate concern among some scholars focusing on the study of intelligence is that any attempt to 'operationalize' critical theory to present-day intelligence analysis could be like trying to shoehorn an interpretivist epistemology into institutional processes that cannot realistically accommodate it.[7] The argument is that any methodology informed by this kind of critical theorizing would result in organizational defensiveness. Critics might ask whether intelligence organizations would allow for their preferred meanings to be deconstructed in an effective way. As 'telling truth' to power provides ample challenges, discussing *truths* would be even more difficult in practice. However, this would greatly depend on the specific context situating intelligence and policymaking. For example, in the United Kingdom there is much more collaboration between intelligence and policy than in the much larger, more competitive and stove-piped US system.

Moreover, rather than dictating 'final assessments' on narratives to intelligence consumers, ACN is much more about cooperation between analysts and working-level policymakers to increase understanding and sensemaking through greater diversity of perspectives. In effect, the ACN approach might serve intelligence consumers more indirectly by informing their staff and subordinates as part of ongoing collaboration. Perhaps what is most necessary is to apprise intelligence consumers of how narratives partly reflect reality, but also shape it, with regard to complex intelligence problems. It is more about consumers and intelligence leadership allowing and facilitating the working-level cooperation necessary for ACN. The integrative approach would eventually also be reflected in finalized intelligence products, and would potentially open up discussions to include policy advice or recommendations for decision making by intelligence consumers. This chapter (and the related research) does not pretend to comprehensively cover how critical theories can contribute to organizational reform.[8] Developing the methodology of ACN itself is not limited by problems that could arise when implementing it in intelligence organizations.

[7] A point raised by Hamilton Bean to the author.

[8] Or other 'projects' on intelligence as, for example, defined by Wesley K. Wark in Ref. [85].

13.4 CONCLUSION

This chapter responds to the dominant positivist paradigm in intelligence studies by outlining and advocating the usefulness of a critical approach. It challenges the so-called invalidity of critical approaches for both the study of intelligence *and* its practice. What distinguishes the ACN approach introduced in this chapter, apart from drawing on different (critical) theory, is its effort to develop a methodology of use for the practice of intelligence that *integrally* analyzes narratives of intelligence consumers, adversaries, and other relevant entities. The theoretical components of critical discourse analysis and securitization theory are useful to infer from theorizing such a 'hybrid' practice of intelligence analysis. ACN is explicit about the necessity of a cooperative working-level effort that brings together different types of officials, including individuals from outside the intelligence organization. They all bring different knowledge and skills to the table, making them the situated interpreters required to analyze a particular narrative. The dialogue in which narratives are contrasted reflects both the necessity to analyze the multi-consequentiality of policy statements and actions, and the need to be self-reflexive about the performativity of the modelling that intelligence analysis encompasses. ACN is also distinct from approaches in both academia and the intelligence practice in its comparative and parallel analysis of threat articulations in at least *three* distinct narratives, two macro (self and other) and one micro (a 'commentator'). Situating and underlying the dynamics in the narratives are fundamentally different social orders. In the contemporary information environment, comprehensively mapping narratives in parallel over time has great value as a primary semiotic mode of entry to complex intelligence problems.

The notion of causal complexes enables us to circumvent the obstacles of classifying securitization as 'successful' and dealing with audiences in terms of causal determinacy. Instead, securitizing actors and audiences are to be situated as causally adequate elements, thus including but also decentralizing the role of audiences for securitization. The study of securitization *efforts* widens the scope on what regular policy practices could be considered to be more indirectly of influence, and hence a relevant part of narratives. The use of the securitization concept in such a manner provides ways to discuss and grasp the nature, status and role of a greater number and more varied types of (potential) audiences. This will show how the critical ACN methodology can potentially contribute to intelligence practice and inform the study of intelligence with new insights. There is also a possibility to contribute to the academic debate on securitization as part of security studies, thus not only learning from, but also linking with other relevant IR subfields.

A more detailed narrative analysis framework and method to trace multi-consequentiality of securitization efforts across narratives, applied in multiple case studies concerning an adequate object of research have been defined and applied by the author elsewhere.[9] That research has successfully analyzed multi-consequentiality of securitization efforts across social domains. Further research is necessary to refine the ACN methodology, and test the practicality of the framework for either *ex post* academic research, or *ex durante* narrative analysis by intelligence analysts, working-level policymakers, and potentially some trusted outside experts, to provide *ex ante* insights for the development of new policies. Following on the initial progress of SAS-114, a successive SAS working group focusing on structured analytic techniques would provide a stimulating and facilitating environment to further the endeavor of developing ACN.

13.5 REFERENCES

[1] De Werd, P. (2018). Critical intelligence studies? A contribution. *Journal of European and American Intelligence Studies,* 1(1):109-148.

[2] Knightley, P. (1986). *The Second Oldest Profession: Spies and Spying in the Twentieth Century*. New York, NY: W.W. Norton & Company.

[9] The author's Ph.D. research includes such a narrative analysis framework and three case studies that demonstrate the ACN methodology.

[3] Betts, R.K. (2007). *Enemies of Intelligence, Knowledge and Power in American National Security.* New York, NY: Columbia University Press.

[4] Kent, S. (1949). *Strategic Intelligence for American World Policy*. Princeton, NJ: Princeton University Press.

[5] Hilsman, R. (1952). Intelligence and policy-making in foreign affairs. *World Politics*, 5(1):1-45.

[6] Platt, W. (1957). *Strategic Intelligence Production: Basic Principles.* New York, NY: F.A. Praeger.

[7] Marrin, S. (2011). *Improving Intelligence Analysis*, 25-28. London, UK: Routledge.

[8] Agrell, W., and Treverton, G. (2015). *National Intelligence and Science: Beyond the Great Divide in Analysis and Policy*, 85-87. Oxford, UK: Oxford University Press.

[9] Kent, S. (1964). Words of estimative probability. *Studies in Intelligence,* Fall 1964:49-65.

[10] Singer, J.D. (1958). Threat-perception and the armament-tension dilemma. *The Journal of Conflict,* 2(1):94.

[11] Gill, P., and Phythian, M. (2012). *Intelligence in an Insecure World*, (2nd ed.), 33-34. Cambridge, UK: Polity.

[12] Lowenthal, M. (2012). *Intelligence: From Secrets to Policy*, (5th ed.),158. London, UK: Sage.

[13] Simms, J. (2014). The theory and philosophy of intelligence. In: *Routledge Companion to Intelligence Studies,* Dover, R., Goodman, M.S., and Hillebrand, C. (Eds.), 42-49. New York, NY: Routledge.

[14] Bhaskar, R., and Hartwig, M. (2008). *The Formation of Critical Realism: A Personal Perspective.* London, UK: Routledge.

[15] Joseph, J. (2012). *The Social in the Global.* Cambridge, UK: Cambridge University Press.

[16] Kurki, M. (2008). *Causation in International Relations: Reclaiming Causal Analysis*. Cambridge Studies in International Relations, (Kindle ed.), Cambridge, UK: Cambridge University Press.

[17] Jessop, R. (2007). *State Power.* Cambridge, UK: Polity.

[18] Wright, C. (2006). *Agents, Structures and International Relations: Politics as Ontology.* Cambridge, UK: Cambridge University Press.

[19] Gill, P. (2010). Theories of intelligence, In *The Oxford Handbook of National Security Intelligence,* Johnson, L.K. (ed.), 43-58. New York, NY: Oxford University Press.

[20] Garrett, D. (2015). *Hume*, The Routledge Philosophers, (1st ed.), London, UK: Routledge.

[21] Fairclough, N., Jessop, R., and Sayer, A. (2010). Critical realism and semiosis. In: *Critical Discourse Analysis,* (2nd ed.), 204. London, UK: Routledge.

[22] Danermark, B., Ekstrom, M., Jakobsen, L., and Karlsson, J.C. (2002). *Explaining Society: Critical Realism in the Social Sciences*, 199. New York, NY: Routledge.

[23] Balzacq, T. (2011). *Securitization Theory*: *How Security Problems Emerge and Dissolve*, New York, NY: Routledge.

[24] Patomaki, H. (2003). *After International Relations*: *Critical Realism and the (Re)Construction of World Politics,* pp. 78-79. London, UK: Routledge.

[25] Patomaki, H., and Wright, C. (2000). After post-positivism? The promises of critical realism, *International Studies Quarterly,* 44(1):213-237.

[26] Oliveira, G.C. (2018). The causal power of securitisation: An inquiry into the explanatory status of securitisation theory illustrated by the case of Somali piracy, *Review of International Studies*, 44(3):1-22.

[27] Wendt, A. (2003). Why a world state is inevitable. *European Journal of International Relations*, 9(4):491-542.

[28] Groff, R. (2004). *Critical Realism, Post-positivism and the Possibility of Knowledge*. London, UK: Routledge.

[29] Furlong, P., and Marsh, D. (2002). A skin not a sweater: Ontology and epistemology in political science, In *Theory and Methods in Political Science*, Marsh, D., Stoker, G. (eds.), 205. Basingstoke, UK, Palgrave Macmillan.

[30] Laclau, E., and Mouffe, C. (1985). *Hegemony and Socialist Strategy: Towards a Radical Democratic Politics*, p. 108. London, UK: Verso.

[31] Montesano Montessori, N., Schuman, H., and de Lange, R. (2012). *Kritische Discoursanalyse: De machten kracht van taal en tekst*, p. 177. Brussels, Belgium: Academic and Scientific Publishers.

[32] Fairclough, N. (2010). *Critical Discourse Analysis,* (2nd ed.) London, UK: Routledge.

[33] Van Dijk, T.A. (2009). Critical discourse studies: A sociocognitive approach. In: *Methods of Critical Discourse Analysis*, Wodak, R., and Meyer, M. (Eds.), 62-86. London, UK: Sage.

[34] Reisigl, M., and Wodak, R. (2009). The discourse-historical approach (DHA). In: *Methods of Critical Discourse Analysis*, Wodak, R., and Meyer, M. (Eds.), 87-121. London, UK: Sage.

[35] Fairclough, N. (1995). *Critical Discourse Analysis, the Critical Study of Language*, (2nd ed.), Boston, MT: Addison-Wesley.

[36] Fairclough, N. (2003). *Analyzing Discourse: Textual Analysis for Social Research.* London, UK: Routledge.

[37] Titscher, S., Meyer, M., Wodak, R., and Vetter, E. (2000). *Methods of Text and Discourse Analysis*, 125. London, UK: Sage.

[38] Somers, M.R. (1994). The narrative constitution of identity: A relational and network approach. *Theory and Society* 23 (5):605-649.

[39] Montesano Montessori, N. (2009). *A discursive analysis of a struggle for Hegemony in Mexico: The zapatista movement versus President Salinas de Gortari*, 147-148. Riga, LV: VDM Verlag.

[40] Fairclough, N. (2015). *Language and Power*, (3rd ed.), New York, NY: Routledge.

[41] Krause, K., and Williams, M.C. (2016). *Critical Security* Studies: Concepts and Strategies, New York, NY. Florence, ITL: Routledge.

[42] Buzan, B., Wæver, O., and de Wilde, J. (1998). *Security: A New Framework for Analysis*. London, UK: Lynne Rienner.

[43] Wæver, O. (2011). Politics, security, theory. *Security Dialogue,* 42(4-5):465-480.

[44] Balzacq, T., Trombetta, M., Stritzel, H., Sjöstedt, R., Schmittchen, D., Vuori, J., Williams, M., Léonard, S., Kaunert, C., Vultee, F., Wilkinson, C., and Salter, M. (2015). In *Contesting Security*, Balzacq, T. (Ed.), New York, NY: Routledge.

[45] Donnelly, F. (2013). *Securitization and the Iraq War: The Rules of Engagement in World Politics*, 49. New York, NY: Routledge.

[46] Stritzel, H. (2007). Towards a theory of securitization: Copenhagen and beyond. *European Journal of International Relations,* 13(3):375-383.

[47] Balzacq, T. (2005). The three faces of securitization: Political agency, audience and context. *European Journal of International Relations,* 11(2):171-201.

[48] Mey, J.L. (2001). *Pragmatics: An Introduction*, (2nd ed.), 221. Oxford, UK: Blackwell.

[49] Balzacq, T., Léonard, S., and Ruzicka, J. 'Securitization' revisited: Theory and cases, *International Relations*, 30(2016)4: 494.

[50] Roe, P. (2008). Actor, audience(s) and emergency measures: Securitization and the UK's decision to invade Iraq. *Security Dialogue,* 39(6):615-635.

[51] Vuori, J. (2008). Illocutionary logic and strands of securitization: Applying the theory of securitization to the study of non-democratic political orders. *European Journal of International Relations,* 14(1):65-99.

[52] Léonard, S., and Kaunert, C. (2011). Reconceptualizing the audience in securitization theory. In: *Securitization Theory,* Balzacq, T. (Ed.), 57-76. New York, NY: Routledge.

[53] Côté, A. (2016). Agents without agency: Assessing the role of the audience in securitization theory. *Security Dialogue*, 47(6):541-558.

[54] Jarvis, L., and Legrand, T. (2017). 'I am somewhat puzzled': Questions, audiences and securitization in the proscription of terrorist organizations. *Security Dialogue,* 48(2):149-167.

[55] Wetherell, M. (2001). Debates in discourse research. In: *Discourse Theory and Practice: A Reader*, Wetherell, M., Taylor, S., and Yates, S. (Eds.), 380-399. Thousand Oaks, CA: Sage.

[56] Salter, M. (2008). Securitization and desecuritization: A dramaturgical analysis of the Canadian Air Transport Security Authority. *Journal of International Relations and Development,* 11(4):321-349.

[57] Salter, M. (2011). When securitization fails: The hard case of counter-terrorism programs. In: *Securitization Theory*, Balzacq, T. (Ed.)., 116-132. New York, NY: Routledge.

[58] Goffman, E. (1959). *The Presentation of the Self in Everyday Life*. New York, NY: Doubleday.

[59] McDonald, M. (2015). Contesting border security: Emancipation and asylum in the Australian context. In: *Contesting Security*, Balzacq, T. (Ed.), 158-159. New York, NY: Routledge.

[60] Foucault, M. (1980). *Power Knowledge: Selected Interviews 1972 – 1977*. Gordon, C. (Ed., Trans.), 194. New York, NY: Pantheon Books.

[61] Jultila, M. (2006). Desecuritizing minority rights: Against determinism. *Security Dialogue*, 37(2):167-185.

[62] Marrin, S. (2017). Why strategic intelligence analysis has limited influence on American foreign policy. *Intelligence and National Security,* 32(6):725-742.

[63] Stritzel, H., and Chang, S.C. (2015). Securitization and counter-securitization in Afghanistan. *Security Dialogue,* 46(6): 548-567.

[64] Hansen, L. (2006). *Security as Practice: Discourse Analysis and the Bosnian War*. New York, NY: Routledge.

[65] Herridge, C. House Intelligence chair: Benghazi attack "Al Qaeda-led event." Fox News, December 29, 2013. Retrieved from http://www.foxnews.com/politics/2013/12/29/house-intelligence-chair-benghazi-attack-al-qaeda-led-event/ (January 10, 2015).

[66] CBS News. 'Face the Nation' transcripts, September 16, 2012: Libyan Pres. Magariaf, Amb. Rice and Sen. McCain. Retrieved from http://www.cbsnews.com/news/face-the-nation-transcripts-september-16-2012-libyan-pres-magariaf-amb-rice-and-sen-mccain/ (January 10, 2015).

[67] Rodham Clinton, H. Remarks on the deaths of American personnel in Benghazi, Libya. September 12, 2012. Retrieved from http://www.state.gov/secretary/rm/2012/09/197654.htm (January 10, 2015).

[68] Gearem, A., and Lynch, C. U.S. Ambassador Susan Rice. *The Washington Post*, October 15, 2012. Retrieved from http://www.washingtonpost.com/world/national-security/us-ambassador-susanrice/2012/10/15/c5a9fe04-16d9-11e2-8792-cf5305eddf60_story.html (January 10, 2015).

[69] Kirkpatrick, D.D. A deadly mix in Benghazi. *The New York Times*, December 28, 2013. Retrieved from http://www.nytimes.com/projects/2013/benghazi/#/?chapt=0 (January 10, 2015).

[70] Simpson, E. (2013). *War from the Ground Up: Twenty-First-Century Combat as Politics*, (2nd ed.), London, UK: Hurst & Company.

[71] Davies, J. (1997). Central Intelligence Agency, A Compendium of Analytic Tradecraft Notes. Central Intelligence Agency, Directorate of Intelligence. Washington, DC. http://www.oss.net/dynamaster/file_archive/040319/cb27cc09c84d056b66616b4da5c02a4d/OSS2000-01-23.pdf

[72] University of Foreign Military and Cultural Studies. (2015). A Tradecraft Primer. In *The Applied Critical Thinking Handbook 7.0*. TRADOC G2, Ft Leavenworth, KS UFMCS. http://nsiteam.com/social/wp-content/uploads/2017/01/RTHB_v7.0_Web.pdf

[73] Chang, W., Berdini, E., Mandel, D.R., and Tetlock, P.E. (2018). Restructuring structured analytic techniques, *Intelligence and National Security*, 33(3): 337-356.

[74] Jones, N. (2018). Critical epistemology for analysis of competing hypotheses. *Intelligence and National Security,* 33(2):273-289.

[75] Coulthart, S.J. (2017). An evidence-based evaluation of 12 core structured analytic techniques. *International Journal of Intelligence and Counter-Intelligence*, 30(2):368-391.

[76] Artner, S., Girven, R.S., and Bruce, J.B. (2016). *Assessing the Value of Structured Analytic Techniques*. Washington, DC: RAND.

[77] Mitchell, G.R. (2006). 'Team B Intelligence Coups'. *Quarterly Journal of Speech*, 92(2):144-173.

[78] Heuer, R., Jr., and Pherson, R.H. (2010). *Structured Analytic Techniques for Intelligence Analysis*, 243. Washington, DC: CQ Press.

[79] Beebe, S.M., and Pherson, R.H. (2012). *Cases in Intelligence Analysis: Structured Analytic Techniques in Action*, Washington DC: CQ Press.

[80] Pherson, K.H., and Pherson, R.H. (2013). *Critical Thinking for Strategic Intelligence.* Washington DC: CQ Press.

[81] Moore, B., and Parker, R. (2009). *Critical Thinking*. 9th edition, New York, NY McGraw-Hill. http://fdjpkc.fudan.edu.cn/_upload/article/10/90/88bc33024683a80cd4da88fc41f0/b44b90d4-3455-4eca-9e9c-31011233c3d1.pdf.

[82] Heuer, R., Jr. (1999). *The Psychology of Intelligence Analysis*, Washington, DC: Center for the Study of Intelligence. https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/psychology-of-intelligence-analysis/PsychofIntelNew.pdf.

[83] Mitchell, W.L., and Clark, R.M. (2015). *Target-Centric Network Modeling: Case Studies in Analyzing Complex Intelligence Issues,* 27-29. Washington, DC: CQ Press.

[84] Hendrickson, N. (2008). Critical thinking in intelligence analysis. *International Journal of Intelligence and CounterIntelligence,* 21(4):679-693.

[85] Wark, W.K. (1993). The study of espionage: Past, present, future? *Intelligence and National Security*, 8(3):1-13.

Chapter 14 – DEDUCTIVE MULTI-VALUED LOGICS FOR PRACTICAL REASONING

Keith K. Niall
Defence Research and Development Canada
CANADA

14.1 REASONING UNDER UNCERTAINTY

How well do people reason in uncertain conditions? There is an approach to the topic which comes from classical logic, meaning systems of deduction. Of course, 'uncertain' may have different interpretations: it may indicate that a proposition is neither true nor false, or it may indicate that a degree of probability is attached to a proposition. Nonsensical or ungrammatical propositions may be neither true nor false, but in this context, uncertainty is meant differently. Within a branch of deductive logic, such uncertainty has been codified as a third truth value alongside 'true' and 'false'. As it is relatively new, different names have been tried for the third value: among them 'uncertain', 'unknown', 'unstatable', and 'indeterminate'. A panoply of deductive systems have been developed that include a third value: they range from elementary three-valued systems to 'fuzzy logic', which has indefinitely many values [1], [2]. Two principal traditions have addressed reasoning in uncertain conditions; one has developed deductive logics in a formal way, while the other has sought to develop a logic of induction. There are different reasons a statement can fall short of being true (or false). A change of state is one reason: what was true may be false later. There exist systems of rules – calculi, if you will – that encapsulate such properties in deductive terms. They evince how less than well-determined states, or how changes in state over time, may be admitted to practical reasoning. This is not to say that rules for practical reasoning must describe how people do reason, or else how they should reason. These systems of rules provide a few standards – maybe unaccustomed standards – for practical reasoning. As nets of logic, they may or may not suit the size of fish we might like to catch in psychological inquiry. The present article proposes multi-valued systems of logic as standards for the investigation of practical reasoning.

A number of issues arise in the treatment of information by intelligence analysts [3]. Two related issues are: the weighting of decisions about possible states of affairs by separate analysts, and decisions about the truth of hypotheses about present events from past information. The arrangement of questions for the analysts, and the combination of their responses, can be considered as a problem in logic. The basic theme of this application of logic is: how do we reconcile simple truth and falsehood with judgements of uncertainty, and how do we reconcile those items we consider true with those we consider contingent over time? The application of multi-valued logic provides the framework for some answers. In **two-valued** logic, propositions are either *true* or *false*: there is no third value. In **fuzzy** logic, there are indefinitely many shades of probability between one and zero [1], [4]. Analysts should not need to make a separate judgement of probability (as in fuzzy logic) apart from considering facts at hand; one can well question the reliability of such judgements of probability, which are likely to be less valuable than assessment of the facts themselves. Multi-valued logic provides a seamless way to combine judgements, and to bridge truth, uncertainty and temporal change. Multi-valued logics exist between classical and fuzzy logics; their interpretation can be framed in terms of possibility (a notion in *modal* logic). The present article sets out a transparent method of combining judgements for the analyst, in the form of truth tables for 3-valued and 6-valued logics and their interpretation.

14.1.1 Issues with Induction

It can hardly be said that the application of deductive logics to practical reasoning is new. Though it may not be new, it is at least unfashionable compared to applications of enumerative induction, and frequentist interpretations of probability. Psychologists are more interested in the growth of knowledge as an

aggregation or an accumulation of knowledge, in the spirit of British Empiricism. It seems to have escaped their notice that "induction, i.e., inference based on many observations, is a myth. It is neither a psychological fact, nor a fact of ordinary life, nor one of scientific procedure ... None of this is altered in the least if we say that induction makes theories only probable rather than certain." (see Ref. [5], p. 53). Part of the trouble with inductive reasoning is that both ordinary knowledge and scientific reasoning progress by **disconfirmation** – they progress in many ways, and in that way also. The elimination of hypotheses, not only their creation or their accumulation, is part of the rational business of experimentation as it is part of empirical discovery.

Carnap sought to develop the logic of induction in a way that presupposes deductive logic. Truth is the first casualty in development of a logic of induction: "all inductive reasoning, in the sense of non-deductive or non-demonstrative reasoning, is reasoning in terms of probability" (see Ref. [6], p. v). Propositions are no longer true and no longer false, only probable. Probability is the second casualty in development of a logic of induction. Probability is split in two: "we have to distinguish chiefly two concepts of probability; the one is defined chiefly in terms of frequency and is applied empirically, the other is a logical concept and is the same as the degree of confirmation". The former notion of probability is not a part of the subsequent development. Inductive logic is not concerned with truth, as are systems of deduction; inductive logic is not concerned with frequentist statistics, either. Induction may be considered the means by which a given proposition is confirmed by prior probabilities. But that statement flows much too easily: prior probabilities may not be relevant to the given one. Some prior propositions bear explanatory value for a proposition, and others do not. Their explanatory value is unrelated to probabilities which may be attached to them. Some conditions do interfere with the confirmation of one proposition by another, as Gettier's examples show [7]. An even more general question can be posed for the confirmation of one proposition by others, and that is: what counts as close? Probabilities may stand proxy for a relation of relevance in probabilist accounts of induction, however often we are reminded that close only counts in horseshoes and in slow dancing. What might be tacit is that the whole program for development of an inductive logic is unfinished – worse, unachieved. A new logic and a new statistics are needed in the search for a logic of induction, but they have scarcely been imagined. It does not help to call them 'mental' logic and 'mental' statistics: that only raises a dust. Even in the theory of inductive logic, application of Bayes's theorem has routinely assumed properties of prior probability distributions [8], and yet "In the classical theory this was done with the help of the principle of indifference. However, since this principle leads to contradictions, we have to give it up" (see Ref. [6], p. 332). Everything is changed here, including the interpretation of Bayes's theorem. Attempts to explain hypothesis formation in science will continue, and the explanation of Baconian induction may be a beautiful, alluring ideal – but it ought not to be paraded as logic.

The application of deductive logics to practical reasoning may seem controversial in psychology, for reasons of history. Philosophy is the birthplace of logic, of course. Experimental psychology separated from philosophy late in the nineteenth century, or one might say early in the twentieth when people still spoke of 'experimental philosophy'. It was more of an acrimonious divorce than it was a gentle separation [9]. In psychology, we may hold a system of logic as a pattern or an ideal for rational argumentation in a domain without insisting that the system constitutes a psychology of reasoning. The separation of psychology from the development of philosophical logic may help to explain the current popularity of induction in psychological explanation. Wayward applications of inductive logic in psychology have developed separately from advances in philosophical logic, as a matter of history.

What is the problem with induction? Accounts of induction have changed since Hume, but induction is still unsatisfactory as logic. "The original difficulty about induction arose from the recognition that anything may follow upon anything" (see Ref. [10], p. 81). Nelson Goodman reveals a major difficulty to a theory of confirmation by induction, which difficulty he calls 'the new riddle of induction' (see Ref. [10], p. 80). Though the difficulties of induction may have changed, "we are left once again with the intolerable result that anything confirms anything. The difficulty cannot be set aside as an annoying detail to be taken care of in due course" (see Ref. [10], p. 75). Like many puzzles in logic, it requires some patience and sensitivity to

appreciate Goodman's new riddle of induction. The problem of induction is the problem of projectibility, meaning "the general problem of proceeding from a given set of cases to a larger set" (see Ref. [10], p. 57). Consider an habitual tawdry example: the proposition *All swans are white*. After one sees a succession of white swans, it might be said that evidence accrues to the proposition, and as a consequence the proposition is more likely. Even Hume did not believe in this sort of induction, when he says: "there can be no **demonstrative** arguments to prove, **that those instances, of which we have had no experience, resemble those, of which we have had experience**" [11]. The problem is not simply that the proposition fails where many swans are black, as in Australia. Perhaps the very notion of induction as enumeration is at fault. Goodman introduced the new riddle to illustrate how traditional enumerative accounts of induction miss the mark of confirmation. Such examples of ambiguity due to change in state over time may seem contingent or unlawlike; their generalization might seem to depend on arbitrary markers. But traditional accounts of induction suppose there can be no markers; that is, they presuppose which regularities will count as lawlike or projectible – and that is the point of the new riddle of induction. It only begs the question to say that one of these predicates and not the other is either primitive or qualitative. Goodman's problem is close to another, more virulent objection concerning the psychological notion of 'following a rule', an objection familiar to readers of Wittgenstein. Kripke says: "I personally suspect that serious consideration of Goodman's problem, as he formulates it, may prove impossible without consideration of Wittgenstein's" (see Ref. [12], p. 59). Induction will not help us in psychological investigation unless we know in advance which predicates will count as projectible: there is no reasonable account of accretion or adduction or aggregation or anything of the kind for the confirmation of knowledge by means of induction. Still many psychologists claim that "scientific progress is a cumulative process of uncertainty reduction ..." (see Ref. [13], p. 349). "But publications on confirmation not only have failed to make clear the distinction between confirmable and non-confirmable statements, but show little recognition that such a problem exists" (see Ref. [10], p. 29). Neither does loose appeal to suspect rules or theorems improve matters. Van Fraassen is no less harsh when he says: "The conclusion is inescapable. Reliability of Induction cannot be part of our scientific knowledge" (see Ref. [14], p. 254).

14.2 DEDUCTION

Deductive methods may be worth a try in the psychological study of reasoning, by contrast. Validity of inference is much better known in systems of deductive logic. Propositions are either true or false in classical two-valued logic. There is no in-between value, by the law of excluded middle – no 'undecided' or 'unknown' as a truth value. Some arguments count as valid, others as invalid. *Modus ponens* and *modus tollens* are valid forms of argument in propositional logic, among others. That is to say, *modus ponens* and *modus tollens* are valid rules of inference in propositional logic. So, for *modus ponens*,

	If this triangle is equilateral, then this triangle is equiangular;	**(A → B)**
	This triangle is equilateral;	**(A)**
Therefore,	*This triangle is equiangular.*	**(B)**

And for *modus tollens*,

	If this triangle is equilateral, then this triangle is equiangular;	**(A → B)**
	This triangle is not equiangular;	**(~B)**
Therefore,	*This triangle is not equilateral.*	**(~A)**

If the premise **(A → B)** is true in the sense of material implication, and premise **(A)** is true, then conclusion **(B)** follows by *modus* ponens. (Table 14-1 shows the truth table for material implication in two-valued logic) If the premise **(A → B)** is true in the sense of material implication, and premise **B** is false.

(~B), then the conclusion **(~A)** follows by *modus tollens*. Both are valid forms of argument, despite whatever response times, preferences, or confidence ratings may be brought to argue that one form of argument may be 'harder' than another [15], [16]. In terms of logic neither is harder: both are valid. (Not unexpectedly for investigators who give pride of place to induction, what appears to be confirmation may be mistaken to be stronger or easier than what appears to be disconfirmation.).

Table 14-1: The Truth Table for Material Implication in a Two-Valued Propositional Calculus. [t: True; *f*: False]. The truth value of a compound proposition P → Q is located in the light gray body of the table, given marginal assignments of truth values, in dark gray, to the propositions P (rows) and Q (columns).

P → Q (material implication)

P \ Q	t	*f*
t	t	*f*
f	t	t

Though an argument is valid, its conclusion may be questioned. Then the premises may be called into question. Consider the following:

	If this triangle is equilateral, then the sum of the internal angles of this triangle is equal to a straight angle (meaning two right angles or 180°);	**(A → C)**
	This triangle is equilateral; and	**(A)**
Therefore,	*The sum of the internal angles of this triangle is equal to a straight angle.*	**(C)**

(A → C) and **(A)** should imply **(C)** by *modus ponens* given that the premises are true. That much *is* true for a plane triangle. But it may be that the sum of the internal angles of a triangle is equal to **three** right triangles for an equilateral triangle on a sphere. **(~C)** is generally true for triangles on a sphere. Suppose that **(~C)** is added to these premises, and replaces **(C).** *Modus ponens* remains valid, but the first premise **(A → C)** does not hold as stated: it is false. If we then choose to infer **(~A,** meaning **A** is not equilateral**)** from the premises **(A → C)** and **(~C)** by *modus tollens*, then both **(A)** and **(~A)** are deemed true. **(A)** was given, and **(~A)** is the conclusion of a valid argument. That is a contradiction, which arises because we have mixed disparate facts and situations. Any proposition at all follows from a contradiction, in two-valued propositional logic.

Acceptance of the validity of a logical argument is not a matter of necessity, meaning: it is neither a logical or a psychological necessity. That is one reason why logic does not serve to endow psychology with laws of thought. It is not a fact of psychology that *modus ponens* or *modus tollens* is a valid argument of propositional logic. Neither is it a law of logic that we are constrained to accept the truth of the conclusion in an argument by *modus ponens*, given the truth of the premises. Lewis Carroll tells a colourful tale to show that *modus ponens* is not a law of logic in that sense [17]. Adherence to the rule is not a compulsion nor a

necessity. As ideals, such rules of inference are **norms of representation**. The point is made in the form of a small fable, 'What the Tortoise told Achilles'. This form of argument adds a hypothetical statement which mentions *modus ponens*. If the hypothetical and the other premises are true, the conclusion follows – but that is just the catch. The conclusion should follow necessarily, given all those premises and another hypothetical statement (Prior notes: "The schoolmen made a distinction here between *necessitas consequentiae*, necessity of the implication, and *necessitas consequentis*, necessity of the implied proposition" (see Ref. [18], p. 115)). The second hypothetical statement is: if the first hypothetical statement and the other premises are true, the conclusion follows. This second hypothetical must be added to the premises, if the conclusion is to be secured as necessary (though such "obtuseness would certainly be phenomenal" (see Ref. [17], pp. 278-280)). This unhappy activity of adding hypotheticals can be extended indefinitely, since there is no end to them. The point of the fable is not that we understand an infinite chain of hypotheticals to apply *modus ponens.* The point is, rather, that we apply *modus ponens* without guarantee of its necessity: it is a norm of representation. No matter how precise the fit between actual judgements and the rules of inference, the latter will not become laws of psychology. They may become useful norms of representation in the study of practical reasoning – but then perhaps not.

14.3 MULTI-VALUED LOGICS

Although systems of deductive logic find a place in mathematical logic (or in what Kleene called 'metamathematics' [19]), not all is settled in the study of logic. Formal systems of modal logic are relatively new, for instance. There have been a few surprises in the modern formalization of classical notions. Every proposition is either true or false in bivalent (two-valued) logic. Some propositions are not well formed in a grammatical sense, but otherwise all propositions count as true or false – including propositions about future states of affairs. Propositions about future states of affairs whose truth is undefined or indeterminate are known as future contingents [20]. Under a strict reading of bivalent logic, in effect there is no room for future contingents. Aristotle knew as much, as did many Scholastic logicians. The treatment of future contingents in bivalent logic bears resemblance to a Calvinist doctrine of predestination in that much, where individuals are eternally saved in advance or eternally damned in advance of an afterlife. The sealing of future contingent propositions seems a rush to judgement. (One would hardly even want to know the truth of propositions about the future, in that case.) Bivalent logic presents some problems for propositions about the past, as well: the causal history of those propositions may be lost, or disjoint, or somehow unattainable. Indeterminacy of propositions about the past can be expressed in **three**-valued logic, with the introduction of a third truth value (for an axiomatic approach, see Ref. [21]).

There are legitimate disagreements on the interpretation of the third truth value; those disagreements lead to distinct formal systems for three-valued logic. Łukasiewicz [22] proposed a three-valued logic, Kleene [19] another, and there are others still [23], [24], [25]. The differences between them can be seen in their truth tables for logical connectives (Table 14-2 and Table 14-4 are truth tables for two logical connectives in Kleene's system, while Table 14-3 and Table 14-5 are truth tables for the analogous connectives in Łukasiewicz's system). Both systems can be generalized (for a generalization of Kleene's system, see Refs. [26], [27]). Łukasiewicz saw his development of a three-valued logic as a modern formal logic necessary to the complete understanding of Aristotle's syllogistic [28]. That is, Łukasiewicz attempted "to set forth modal syllogistic as a completely formalized deductive system" [29], though Łukasiewicz's faithfulness to the Peripatetic system can be questioned in detail (see Ref. [30], pp. 395-404). In this context it is worth noting that "Łukasiewicz seems inclined to interpret all modal logics after the general pattern of his 3-valued logic of 1920, and he says explicitly that all modal logics, in his sense of the term, must be many-valued" (Refs. [31]; [32], pp. 170-171). His was an early intuition of the equivalence of some modal logics and multi-valued logics.

Table 14-2: The Truth Table for Material Implication in a Three-Valued Calculus, Under Kleene's [19] Interpretation of a Third Truth Value. [t: True; u: Undefined; *f*: False].

$P \rightarrow_K Q$ **(material imp. for Kleene)**

P \ Q	t	u	*f*
t	t	u	*f*
u	t	u	u
f	t	t	t

Table 14-3: The Truth Table for Material Implication in a Three-Valued Calculus, Under Łukasiewicz's [22] Interpretation of a Third Truth Value. [t: True; i: Intermediate; *f*: False].

$P \rightarrow_Ł Q$ **(material imp. for Łukasiewicz)**

P \ Q	t	i	*f*
t	t	i	*f*
i	t	t	i
f	t	t	t

Table 14-4: The Truth Table for Weak Equivalence ('if and only if') in a Three-Valued Calculus, Under Kleene's [19] Interpretation of a Third Truth Value. [t: True; u: Undefined; *f*: False].

$P \equiv_K Q$ **(iff for Kleene)**

P \ Q	t	u	*f*
t	t	u	*f*
u	u	u	u
f	f	u	t

Table 14-5: The Truth Table for Weak Equivalence ('If and only If') in a Three-Valued Calculus, under Łukasiewicz's [22] Interpretation of a Third Truth Value. [t: True; i: Intermediate; *f*: False].

$P \equiv_Ł Q$ (iff for Łukasiewicz)

P \ Q	t	i	*f*
t	t	i	*f*
i	i	t	i
f	f	i	t

14.3.1 Tense Logic

There is a second intuitive notion which leads us deeper into multi-valued logic. Yes; some propositions are neither true nor false, but also consider that the truth of some propositions changes with time. That circumstance may be addressed, not by a change in copula within propositions as some have proposed, but by further development of multi-valued logics in the form of **temporal** logic. Lewis Carroll points out the need for a tensed logic in practical reasoning. He poses an invalid argument in the form of a syllogism. The plain-language form of this argument should be valid, which follows [33]:

> *The meat that I eat at dinner is meat that I buy at the market;*
>
> *The meat that I buy in the market is raw meat.*
>
> *Therefore, the meat that I eat at dinner is raw meat.*

The statement of the conclusion would be true at least if I had had steak tartare for dinner, which I don't make very often. The example jars our intuition: either the argument is invalid, or it seems invalid because it ignores a change in state over time. This provides an illustration of the need for a temporal logic, though it is far from the full rationale.

Arthur Prior developed several systems for a logic of tenses; among them is his system ***Q*** [31]. Prior continued to develop his logic of tenses [18], [34], and others have continued his line of investigation [35]. Prior's system connects with other logical systems: many of his temporal logics are modal logics. The temporal distinction which is important in system ***Q*** is that between *today* and *yesterday* (one might say: *today* and *before*). System ***Q*** is capable of handling statements whose truth value is different today from what it was yesterday. Prior's system retains a third truth value besides *true* and *false*. Like Kleene and like Łukasiewicz before him, he wondered how best to express this value in ordinary speech. They had identified it as 'undefined' and 'intermediate' respectively, avoiding epistemological connotation. Prior eventually settled on 'unstatable' as an expression for this third value in system ***Q***. The possible values of statements in ***Q*** are then:

1) True at both times;
2) True today and unstatable yesterday;
3) True today and false yesterday;

4) False today but true yesterday;

5) False today and unstatable yesterday; and

6) False at both times.

We will label these values as: **t**, $\mathbf{i_1}$, $\mathbf{i_2}$, $\mathbf{i_3}$, $\mathbf{i_4}$, and ***f***. The truth tables for three logical connectives in system ***Q*** are given as Table 14-6, Table 14-7, and Table 14-8.

Table 14-6: The Truth Table for 'Exclusive Or' (P V Q) in Prior's [31] Six-Valued Calculus *Q*. [t: True today, true before (designated); i_1: True, unstatable before (designated); i_2: True today, false before; i_3: False today, true before; i_4: False today, unstatable before; *f*: False today, false before].

P \ Q	t	i_1	i_2	i_3	i_4	*f*
t	t	i_1	t	t	i_1	t
i_1	i_1	i_1	i_1	i_1	i_1	i_1
i_2	t	i_1	i_2	t	i_1	i_2
i_3	t	i_1	t	i_3	i_4	i_3
i_4	i_1	i_1	i_1	i_4	i_4	i_4
f	t	i_1	i_2	i_3	i_4	*f*

Table 14-7: The Truth Table for a Two-Place Connective for Implication in Prior's [31] Six-Valued Calculus *Q*. [t: True today, true before (designated); i_1: True unstatable before (designated); i_2: True today, false before; i_3: False today, true before; i_4: False today, unstatable before; *f*: False today, false before].

P \ Q	t	i_1	i_2	i_3	i_4	*f*
t	t	i_1	i_2	i_3	i_4	*f*
i_1	i_1	i_1	i_1	i_4	i_4	i_4
i_2	t	i_1	t	i_3	i_4	i_3
i_3	t	i_1	i_2	t	i_1	i_2
i_4	i_1	i_1	i_1	i_1	i_1	i_1
f	t	i_1	t	t	i_1	t

Table 14-8: The Truth Table for Conjunction 'and' in Prior's [31] Six-Valued Calculus *Q*. [t: True today, true before (designated); i_1: True unstatable before (designated); i_2: True today, false before; i_3: False today, true before; i_4: False, unstatable before; *f*: False today, false before].

P \ Q	t	i_1	i_2	i_3	i_4	*f*
t	t	i_1	i_2	i_3	i_4	*f*
i_1	i_1	i_1	i_1	i_4	i_4	i_4
i_2	i_2	i_1	i_2	*f*	i_4	*f*
i_3	i_3	i_4	*f*	i_3	i_4	*f*
i_4	i_4	i_4	i_4	i_4	i_4	i_4
f	*f*	i_4	*f*	*f*	i_4	*f*

A distinction between *today* and *yesterday* (if you will: *today* and *before*) may seem too simple to account for changes in state over time. Prior is quick to point out a more comprehensive extension of ***Q*** with a greater number of values: "These 6-valued tables are in fact just the first step towards the matrix with an infinite number of elements which would give us the exact many-valued equivalent of the system ***Q***" [18]. System ***Q*** is not only a useful logic of tense; it is also a modal logic in the ordinary sense [18]. Though ***Q*** can be adapted to future contingent statements, as a logic of tense the extended system has a greater range.

14.3.2 Design of an Experiment with Tense Logic

How could system ***Q*** be used in a psychological study? A small example begins with a state of statements – a set of propositions, if you like – about the political geography of North America post-European colonization and present day. (The reader may find that our example privileges some facts over others: say, facts about US political boundaries over the political geography of Mexico. The domain of facts can be expanded to be inclusive. Such examples turn ... well, *political*.) Some readers may want to amplify the face validity or cogency of this small example: they are encouraged to develop parallel examples of the development of political boundaries in the Golan Heights, or of historical changes in the map of Africa. Here are a few propositions about the political geography of North America (see Figure 14-1): all of them have been true at one time (for the purpose of these examples, interpret 'true before' as 'true at least one time in the past'):

- *France cedes Louisiana to Spain.*
- *The province of Canada is divided into the two provinces of Ontario and Québec.*
- *Spain cedes Louisiana to France.*
- *Texas is part of the Confederacy.*
- *Some of the Virgin Islands belong to Denmark.*
- *Québec is divided into Upper and Lower Canada.*

- *Spain cedes Louisiana to France.*
- *Canada extends to the North Pole.*
- *Massachusetts borders Québec and New Brunswick.*
- *The eastern border of Manitoba is disputed by the province of Ontario.*

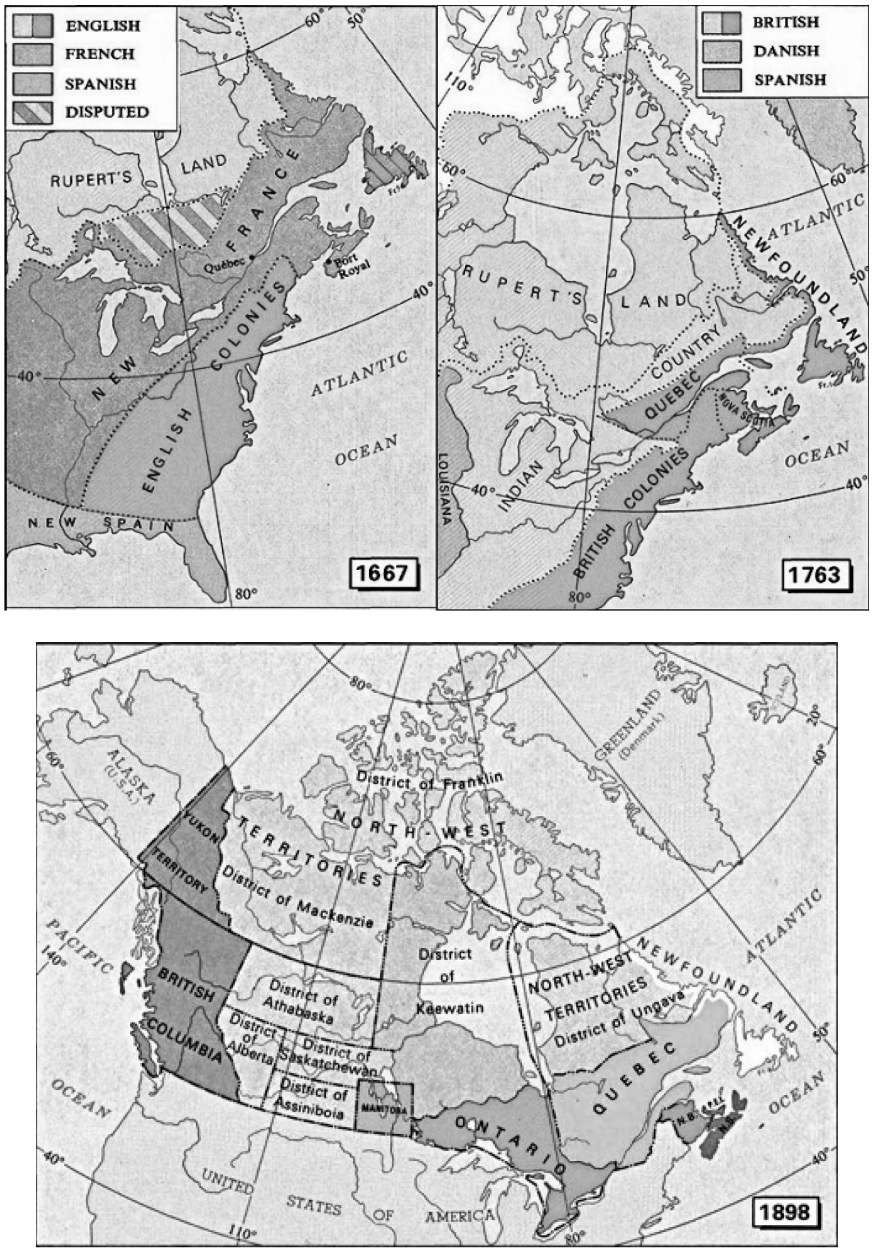


Figure 14-1: Three Historical Maps of Canada's Political Boundaries, from 1667, 1763, and 1898.

Note: The maps are taken from National Resources Canada, selected maps from the *Atlas of Canada* (1957), 3rd edition. Sites accessed April 2017. *Open Government License – Canada: 'You are free to: Copy, modify, publish, translate, adapt, distribute or otherwise use the Information in any medium, mode, or format for any lawful purpose.'*

It is not too controversial to say that such statements can be true, and that some have changed over time from being true to being false. There is enough in these statements to make logicians squirm: entities which have changed name, entities which no longer exist, and more. However, we shall take them as having been true, false, or undecided, either now or before, inasmuch as they are functional for historical and cartographic purposes.

The larger number of statements in our example concern colonial political decisions about regions of North America. This choice of topics could be considered a facile choice, an easy way out. Decisions about political boundaries can be expressed with performative utterances (as in "I claim this land for France"), and in a sense, performative utterances are definite. They are definite given a hospitable matrix of rules and conditions (or else Woodrow Wilson points to a map and says: "This land belongs to Armenia."). They are defeasible and they may be disputed, but they are not often intentionally vague (Austin's exhortation should not be taken as a theory of truth, for which effort he has great respect) [36]. Sometimes the act associated with the utterance is not achieved ("I claim this land for France" says the actor playing Samuel de Champlain), but what's done is done. The point is not so much that our example about political geography is easy, but rather that greater consideration of the many uses of language may be difficult. "Very few statements that we ever utter are just true or just false. Usually there is the question are they fair or are they not fair, are they adequate or not adequate, are they exaggerated or not exaggerated? Are they too rough, or are they perfectly precise, accurate, and so on? 'True' and 'false' are just general labels for a whole dimension of different appraisals which have something or other to do with the relation between what we say and the facts" (see Ref. [36], pp. 237-238). A move away from enumerative induction towards deductive multi-valued logics shouldn't seem an easy victory for the application of logic to the psychology of reasoning, therefore. It is one step in the right direction.

How can we assess how closely sets of judgements align to valid rules of inference for multi-valued logics? Let us take an example for the six-valued logic as our central example, since that arrangement may seem more complex. We may begin with some uncontroversial declarative propositions, which is not to say uncontroversial propositions are easy to find. We can express those propositions in sentences, either singly or as joined in pairs by logical connectives. People are then asked to assign a truth value (one of 2, 3, or 6 values, depending on the logic in question) to the single propositions. They are also asked to assign a truth value (one of 2, 3, or 6) to the compound propositions joined by logical connectives (Figure 14-2 shows how six values may be presented).

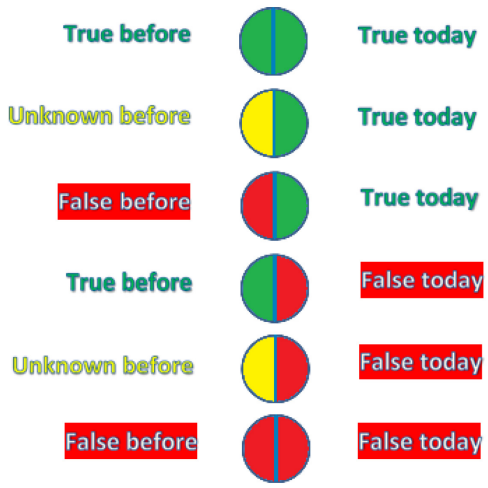


Figure 14-2: Choices Among Truth Values for Each Proposition may be Facilitated by an Easy and Colourful Arrangement of the Alternatives. Here is one suggestion for such an arrangement.

A range of logical connectives can be employed: there are n^{n^2} truth functions of two arguments in an n-valued logic (i.e., 8, 81, and 2, 176, 782, 336 respectively) [37]. We would like to make the task easy by using pairs of propositions rather than, say, long formulae in unfamiliar notation. Here we make a choice for experiment: we shall present the single propositions first in a group, followed by the group of compound propositions. Naturally enough, these single propositions have truth values, and the compound propositions have truth values which follow the truth function of the logical connective in their sentence. Many compound propositions should be presented, with component propositions of different truth values. Enough compound propositions should be presented, that the places in a 'truth table' for the logical connective are mostly filled. For the purpose of later statistical evaluation, enough compound propositions should be presented that the places in the 'truth table' which are filled are filled more than once or twice. Many compound propositions should be presented, with component propositions of different truth values. Enough compound propositions should be presented, that the places in a 'truth table' for the logical connective are mostly filled. For the purpose of later statistical evaluation, enough compound propositions should be presented that the places in the 'truth table' which are filled are filled more than once or twice. Not all the places need contain an identical number of entries, though this will arise later as an empirical question concerning frequency of presentation. What counts as a compound proposition? Consider that Ottawa is now the capital of Canada; Kingston, Ontario was once the capital of Canada (overlooking that Canada was then the province of Canada). Toronto is not the capital, and Toronto has never been the capital of Canada. Then the following four compound propositions are true:

1) Either at no time was Ottawa the capital of Canada **[*f*: False today, false before],** or at one time Kingston was the capital of Canada **[t: True today, true before],** but not both.
2) If Ottawa is the capital of Canada **[i_2: True today, false before]**, then Kingston is not the capital of Canada **[i_2: True today, false before]**.
3) If at one time Toronto was the capital of Canada **[*f*: False today, false before],** then at one time Kingston was the capital of Canada **[t: True today, true before]**.
4) If Kingston is the capital of Canada **[i_3: False today, true before],** then Ottawa is not the capital of Canada **[i_3: False today, true before]**.

What is important in the test (and in the analysis) is not the formation of compound propositions, but the consistency of judgements on the compounds with judgements of their component single propositions. In the empirical test, those single propositions are judged prior to the compound propositions – in one condition with some help, and in the other condition without help. Consider the single proposition Labrador *is a part of Québec*. In one of the conditions some help is given after each single proposition, after an initial judgement is made. The help would be given as additional information, roughly in this manner:

- Labrador is a part of Québec. **[i_3: False now, True before]**.

Year: 1774 The Quebec Act transfers Labrador, Anticosti Island and the Magdalen Islands from Newfoundland to Québec.

Year: 1791 Québec is divided into Upper and Lower Canada.

Year: 1809 Labrador and Anticosti Island are transferred from Lower Canada to Newfoundland.

People will assign a truth value to many single propositions, and to many compound propositions. A number of different logical connectives should be assayed, meaning that people will be asked to assign truth values corresponding to all places in the truth tables for several different connectives. Mostly, two-place connectives will stand proxy for longer and more complex formulae of the logic in question, since presumably propositions containing three- or four-place connectives are more difficult to judge. It is an empirical matter how accurately people may assign truth values with practise, or with training, or else how

consistent their assignments may be to the assignments made by other subjects. All that is important to an empirical test. Yet in a sense, a person's assignments of individual truth values are just the means to an end. The 'truth table' for a logical connective is the **unit of observation** in this investigation. Primarily we are interested in each complete table assembled from the many judgements a person makes. (It may be controversial that some single propositions can be given at first as being previously unknown; the relevant conditions may be omitted altogether, but then the subsequent analysis of errors in judgement will be only partially complete.) We are interested in the adherence or the alignment of that table to the truth table for the two-place logical connective in question (unless we dare for a three-place connective, or more). Particular mistakes are of secondary concern to us, that is, they are of lesser concern compared to the fit that a person's many judgements may provide to a whole table. The person's many judgements serve the end of assaying their interpretation of a logical connective. A person's interpretation of varied logical connectives is the aim of the inquiry, in effect. People do make errors, though, and we can ask how degree of fit to the pattern of a truth table should best be evaluated. Which two-place connectives should be considered? Let us begin with a small familiar set:

$$\mathbf{P} \vee \mathbf{Q};\ \mathbf{P} \rightarrow \mathbf{Q};\ \mathbf{Q} \rightarrow \mathbf{P};\ \text{and}\ \mathbf{P} \,\&\, \mathbf{Q}.$$

These are items which may be incorporated in longer formulae (though there is no psychological license yet to combine them for experimental prediction).

14.3.3 Evaluation of Results on Tense Logic

How can a truth table be compared to a table constructed from a person's judgements? We can establish a confusion matrix from the two tables. This can be done in a couple of ways: by comparing a person's judgements of compound propositions to the truth table for the connective, or by comparing the person's earlier judgements of single propositions to their later judgements of compound propositions – containing those single propositions joined by the two-place connective. In the former case, judgements are expected to coincide with truth values in each place of the truth table. Yet anyone could be ignorant of these propositions. And so, in the latter case, the judgements should coincide with truth values in each place, **supposing** that the person's initial assignments of truth values to single propositions are correct assignments. That is: we compare the judgements to truth values **under the hypothesis** that the person's initial judgements were correct. (The reliability of those initial judgements is another question for experiment.) We can establish a confusion matrix in either case, with the same number of rows and columns as the truth table; we put a **zero** in each place where judgement and truth value fail to coincide, and we put a **one** where they do coincide. (Again, the flavour of people's mistakes is of secondary interest for our purposes.) The values of zero and one accumulate. The **expected value** at each place is determined by the truth value for the connective, and by the frequency of presentation of the propositions. The **observed value** at each place is determined by the number and type of the person's judgements of the compound propositions. The residual at each place is the observed value minus the expected value. On this analogy, confusion matrices and residuals can also be computed to compare one person's judgements with another's.

The entire truth table for a logical connective is the unit of observation for any psychological investigation. Large truth tables in 3 x 3 format or 6 x 6 format may seem awkward for the purpose of experiment. Yet it would be inappropriate to test an individual's performance on component blocks of adjacent truth values within the truth tables, in the expectation that those selected tests could be welded together to represent the whole truth table. Truth tables may be subject to permutation by rows or permutation by columns while retaining their interpretation, and this implies that the adjacency of truth values for a table is accidental. Any operation of 'welding together' component blocks is the imposition of a logical operation: it favors one pattern of judgement over others which might have been chosen. Another issue arises because an individual may confound one logical operation with another when assigning truth values. With ambiguous instructions an individual may be led to confuse an exclusive 'or' with an inclusive 'or', for instance. (Inhelder and Piaget's studies on the development of logical ability have been criticized on just these grounds [38].)

Individuals do make errors in judgement, which eventuality compounds the confusion further. The 2 x 2 component block of an individual's responses may represent or may correspond to many things. Say that an experimenter means the 2x2 block to represent four truth values for material implication in Łukasiewicz's interpretation of that connective. The block of judgements could also be taken as part of the truth table for material implication on Kleene's interpretation. One interpretation can be mistaken for another, just when it is the purpose of an experiment to distinguish them. That is not all. The 2x2 block may differ but little from a similarly situated (in terms of truth values of its component propositions) set of truth values within Kleene's table for the connective 'if and only if'. Or else: the 2x2 block may align well to an analogous part of Łukasiewicz's table for 'if and only if', or else to part of a truth table for yet another connective, and so forth. The surest course for experiment is to retain the entire truth table as one's unit of observation. Individuals may still mistake one logical connective for another, but their errors will be consistent, and uncompounded by any unwarranted imposition of logic by interpretation of their judgements from component tables.

A summary statistic can be calculated on each table of judgements. That summary statistic is the likelihood ratio criterion,

$$\chi^2{}_L = 2\sum[\boldsymbol{O}bserved \times ln\left(\boldsymbol{O}bserved/\boldsymbol{E}xpected\right)] \tag{14-1}$$

which is distributed as χ^2 with (*#rows* – 1)(*#columns* – 1) as its degrees of freedom. (A sign test on the residuals may also be applied as a binomial statistic.) An additional test of the symmetry of these tables can be computed in some instances. The truth tables for **P V Q** and **P & Q** are expected to be square and symmetric across the major diagonal; the tables of residuals associated with them may be tested for adherence to that symmetry. A statistic for this test of symmetry is given by:

$$\boldsymbol{\chi^2} = \sum_{i<j} [(n_{ij} + n_{ji})^2 / (n_{ij} + n_{ji})] \tag{14-2}$$

which is distributed as χ^2 under the hypothesis of symmetry for a square table. That χ^2 statistic has $(\frac{1}{2}\#rows(\#rows-1))$ as its degrees of freedom. "The distribution of χ^2 is asymptotically distribution-free (i.e., free of the influence of the hypothetical distribution's form and parameters) ... " (see Refs. [39]; [40], 443). Our summary statistic may be used in various ways, including a comparison of groups by the analysis of variance. One could compare the effects of prior instruction on the performance of a group of individuals against the performance of another group of individuals who had not received the same instruction. (A difference in instruction may be gauged by the truth values assigned to the group of single propositions administered first, for instance.) The summary statistic for an individual would serve as a unit for the analysis of variance in that case. The statistics of this comparison are transparent, since the ***F*** statistic applied in the analysis of variance is equivalent to a $\boldsymbol{t}^2$ statistic for two groups; also note that the ratio of χ^2 statistics is distributed as ***F***.

14.4 DISCUSSION WITH SUGGESTIONS FOR SOFTWARE

> *Pheidippides (the son): And what is it I should learn?*
>
> *Strepsiades (the father): It seems they have two courses of reasoning, the true and the false, and that, thanks to the false, the worst law-suits can be gained. If then you learn this science, which is false, I shall not have to pay an obolus of all the debts I have contracted on your account.*
>
> *Pheidippides: No, I will not do it. I should no longer dare to look at our gallant horsemen, when I had so ruined my tan*. [41]

A psychology of practical reasoning should aim to say how we reason, meaning how our uses of rules of inference do follow norms of representation for logic. One motivation for developing a psychology of deductive reasoning would be to demonstrate how the implementation of semi-automated assistance can

improve practical reasoning [42]. A psychology of practical reasoning is then not an account of naïve errors in reasoning, or of the biases of the untaught (as in Refs. [43], [44]). Of course, we may not reason perfectly if we lack time or patience or motivation. When presented a set of premises, we do not know or intuit at once every conclusion which may be drawn from them. There are people – intelligence analysts are an example – whose task it is to draw valid conclusions from large or even enormous funds of propositions (as in Ref. [45]). (Not all of their 'information' will be in propositional form, naturally.) Some assistance in reasoning could be provided to them. That might be in the form of the truth value of a premise they invent, as it is added to their existing stock of premises. Or it could be a list of premises in that stock from which a proposition of interest follows, and so on. There is no shame in relying on a little assistance in reasoning – unless we mistake such assistance for an independent intelligence, as for instance an artificial intelligence. One might ask, too: how can complex schemes of multi-valued logic be used in conscious reasoning? Perhaps unexpectedly, one answer is: any demarcation between reasoning which is conscious and unconscious reasoning must have a functional role in the psychology of reasoning, and the psychology of reasoning is inchoate. The meaning and place of such a distinction remains obscure. All human reasoning might well be called unconscious or else conscious because the distinction is applied prematurely: it awaits our better understanding.

How may these systems of logic be mirrored in software as an adjunct to reasoning? The aim should be to facilitate reasoning, not to encourage delusions that human judgement may be supplanted. The instantiation of such systems should enable three features of human reasoning: hypothesis selection, disconfirmation, and synopsis. That is not to succumb to stock-in-trade terms of induction with talk of condensing masses of data; 'Proposition' does not become a mass noun even when there are many of them. Talk of quantities of information or volumes of evidence (i.e., quarts, not folios) is just misplaced metaphor in advance of the application of logical systems. Here, hypothesis selection is not meant in the sense of the generation of fresh propositions. It is meant as a technique of placing propositions **under hypothesis**, by bracketing them from or holding them apart from the context of a larger argument. Then the suppositions and consequences of those propositions may be explored before they are set among the commitments of the larger and more inclusive argument, in other words without affecting truth values in the wider context. In non-technical terms, this is the creation of a scratch-pad or worksheet for subsidiary sets of propositions. In more technical terms, it is a procedure for reasoning under hypothesis, to allow an individual to explore presuppositions of selected propositions [46]. For disconfirmation to be enabled in software, an individual would be entitled to negate (and remove) one or more of the existing hypotheses of a system; that would have the effect of marking (of 'flagging') the consequences of that negated proposition. Disconfirmation has a positive purpose, as pruning has in arboriculture: investigation proceeds when "**we do not seek highly probable theories but explanations; that is to say, powerful and improbable theories**" (see Ref. [47], 58, 216). High-probability propositions say little; they lack explanatory power. Popper's discussion of disconfirmation began with the growth of scientific knowledge, but he adds: "my remarks are applicable without much change, I believe, to the growth of pre-scientific knowledge also – that is to say, to the general way in which men, and even animals, acquire new factual knowledge about the world". Synopsis should be a third feature of adjuncts to reasoning. Synopsis in this sense is not the condensation of content or the truncation of content. It means the provision of a broad overview, a surview of the current domain of propositions. Of the three, this is the most difficult feature to implement well. Two options seem to lie open for its implementation. One is the development of better techniques for the visualization of large connected graphs to represent systems of propositions. (Let us work for the evolution of this visual representation beyond a morass of edges and labelled nodes.) Another is the instantiation of **natural deduction** [48] to work towards simpler and more relevant sets of entailed propositions. "The intuitive idea lying behind systems of natural deduction is there should be, for each logical connective, one rule justifying its Introduction into discourse, and one rule for **using** ('eliminating') the connective once it has been introduced". These are a few suggestions towards an adjunct to reasoning: they are not procedural instructions. Still we have come this far: given a large corpus of propositions, the instantiation of a multi-valued temporal logic affords a useful adjunct to reasoning, if it allows for hypothesis selection, for disconfirmation, and for synopsis.

And in the end, people are bound to ask: what are the products of your exercise? There are a few interesting products which emerge from the present investigation:

1) A method of testing was established to determine how well practical judgements may follow rules of inference. The aim of the test is fairly unobtrusive to the people involved.
2) This method is suited to distinguish between a more frequent 'natural' interpretation of implication in three-valued logic, and a less-frequent interpretation.
3) The application of a six-valued temporal logic, Prior's **Q**, was outlined. That logic facilitates judgements on propositions involving changes of state over time.
4) A preliminary measure of the practical difficulty of rules of inference has been developed.
5) The logical systems that have been applied are extensible, and may be automated in part: diverse categories of uncertainty, and more elaborate temporal logics can be brought to bear as necessary.

The best use of a psychology of practical reasoning is to help us in reasoning; the study of practical reasoning should aim to improve inference in pragmatic situations. A sketch has been made of experiments in psychology; there are some slight details that can be improved. First, the materials must be sound: the logical form of the propositions is to be univocal. The results of such experiments must also be replicable across propositions drawn from widely different subject domains [47], [49]. Then there are moderate improvements to be made by developments in formalism (e.g., Ref. [50]). The extension of Prior's temporal logic to many temporal intervals has been mentioned [18]. Other formalisms of logic can be borrowed from deontic [51], doxastic, and epistemic modes, too, but much stronger stuff is required for a psychology of practical reasoning. Two things are necessary, at least: we need to incorporate yet more advanced logic in our studies, and we need to begin to account for expert knowledge in reasoning. The subject matter of a psychology of practical reasoning is what we do at our best, our reach for the ideal: how experts apply advanced logic. The psychology of practical reasoning begins when we recognize that advanced logic has a place in psychology even if psychology should have no place in logic (if, as Łukasiewicz puts it: "What is called 'psychologism' in logic is a mark of the decay of logic in modern philosophy" (see Ref. [28], 13)). We may also develop more sophisticated goals (A starred exercise for application by the reader: Refs. [52], pp. 67-96; [53], pp. 206-220), including the salutary effects of explicit and extended instruction on logic for occupations whose business it is to draw conclusions from masses of information.

Of course, the final question is: why should we study multi-valued logic – how should it help intelligence analysts? Truth is still a useful notion for analysts; in a sense it is their job to preserve truth. Logic expresses how truth is preserved in valid entailment. Logic preserves **truth**, more than and not simply just by probabilities. Multi-valued logics express validity of entailment under conditions of uncertainty, and under changes of state over time. These logics reveal patterns of inference that formal systems of probability never can, and never will. If the maintenance of truth is worth anything, multi-valued logics set the norm for reasoning for intelligence analysts (and for everyone else) under such conditions. There is a story about the cynicism of Pontius Pilate, the Roman prefect of Judaea who asked "What is truth?", but would not stay for an answer. Pilate was ahead of his time.[1]

14.5 REFERENCES

[1] Zadeh, L.A. (1965). Fuzzy sets. *Information and Control,* 8:338-353.

[2] Vink, M.P., and van Vliet, O. (2009). Not quite crisp, not quite fuzzy? Assessing the potentials and pitfalls of multi-value QCA. *Field Methods,* 21(3):265-289.

[1] This means more than a biblical reference to John 18:38. Francis Bacon used this very quip to open his essay "On truth", which appeared in 1597. John Austin also used it to open his 1961 essay on truth (see Ref. [36]).

[3] Palo Alto Research Center in collaboration with Heuer, R.J., Jr. (2010). *Analysis of Competing Hypotheses: ACH2.0.5*. Palo Alto, CA: PARC. Retrieved from http://competingghypotheses.org (February 2017).

[4] Zadeh, L.A. (1997). Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems,* 90:111-127.

[5] Popper, K. (1965). *Conjectures and Refutations,* (2nd. ed.), London, UK: Routledge and Kegan Paul.

[6] Carnap, R. (1962). *Logical Foundations of Probability*, (2nd ed.), Chicago and London: The University of Chicago Press.

[7] Gettier, E. (1963). Is justified true belief knowledge? *Analysis* 23, (6):121-123.

[8] Bayes, T. (1764). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* 53:370-418.

[9] Macnamara, J. (1986). *A Border Dispute: The Place of Logic in Psychology*. Cambridge, MA: MIT Press.

[10] Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

[11] Hume, D. (1978, originally published 1739). *A Treatise of Human Nature. Vol. 1: Of the Understanding*. Oxford: The Clarendon Press. See Book I, Section 5, 89.

[12] Kripke, S. (1982). *Wittgenstein on Rules and Private Language: An Elementary Exposition*. Cambridge, MA: Harvard University Press.

[13] Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science,* 349 (6251):aac4716-7.

[14] Van Fraassen, B.C. (2000). The false hopes of traditional epistemology. *Philosophy and Phenomenological Research,* 60(2):253-280.

[15] Alós-Ferrer, C., Garagnani, M., and Hügelschäfer, S. (2016). Cognitive reflection, decision biases, and response times. *Frontiers in Psychology,* 7(1402):21.

[16] Nakamura, H., and Kawaguchi, J. (2016). People like logical truth: Testing the intuitive detection of logical value in basic propositions. *PLoS ONE,* 11(12):e0169166.

[17] Carroll, L., pseudonym of Dodgson, C. (1895). What the tortoise said to Achilles. *Mind*, 4(14):278-280.

[18] Prior, A.N. (1967). *Past, Present and Future*. Oxford at the Clarendon Press.

[19] Kleene, S.C. (1952). *Introduction to Metamathematics*. New York and Toronto: D. Van Nostrand Company, Chapter XII.

[20] Prior, A.N. (1953). Three-valued logic and future contingents. *The Philosophical Quarterly*, 3(13):317-326.

[21] Rose, A., and Rosser, J.B. (1958). Fragments of many-valued statement calculi. *Transactions of the American Mathematical Society,* 87:1-53.

[22] Łukasiewicz, J. (1930). Untersuchungen über den Aussagenkalkül. *Comptes rendus des séances de la Société des Sciences et des Lettres de Varsovie*, Classe III, 23:30-50.

[23] Parks, R. Zane. (1972). A note on R-mingle and Sobociński's three-valued logic. *Notre Dame Journal of Formal Logic,* 13(2):227-228.

[24] Figallo, A., and Ziliani, A. (1992). On the propositional system A of Sobociński. *Portugaliae Mathematica*, 49(1):11-22.

[25] Ciucci, D., and Dubois, D. (2013). A map of dependencies among three-valued logics. *Information Sciences,* 250:162-177.

[26] Fitting, M. (1992). Kleene's logic, generalized. *Journal of Logic and Computation,* 1:797-810.

[27] Fitting, M. (1994). Kleene's three-valued logics and their children. *Fundamenta Informaticae*, 20:113-131.

[28] Łukasiewicz, J. (1951). *Aristotle's Syllogistic, from the Standpoint of Modern Formal Logic*. Oxford: The Clarendon Press.

[29] McCall, S. (1963). *Aristotle's Modal Syllogisms*. Amsterdam, Netherlands: North-Holland.

[30] Austin, J.L. (1952). Critical notice for Aristotle's syllogistic. *Mind*, 61(243):395-404.

[31] Prior, A.N. (1957). *Time and Modality*. Oxford: The Clarendon Press.

[32] Łukasiewicz, J. (1920). O logice trójwartościowej. [On three-valued logic]. *Ruch Filozoficzny*, 5:170-171.

[33] Bartley, W.W., III. (1977). *Lewis Carroll's Symbolic Logic*. New York, NY: Clarkson N. Potter, Inc. Logical puzzle, § 6, 426.

[34] Prior, A.N. (1968). *Papers on Time and Tense*. Oxford at the Clarendon Press.

[35] Copeland, B.J. (Ed.). (1996). *Logic and Reality: Essays on the Legacy of Arthur Prior*. Oxford at the Clarendon Press.

[36] Austin, J.L. (1961). *Philosophical Papers*. Oxford at the Clarendon Press.

[37] Haack, S. (1978). *Philosophy of Logics*, 30. Cambridge, UK: Cambridge University Press.

[38] Inhelder, B., and Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence*. London, UK: Routledge and Kegan Paul.

[39] Everitt, B.S. (1992). *The Analysis of Contingency Tables*, (2nd ed.) London, UK: Chapman and Hall.

[40] Kendall, M.G. (1967). *The Advanced Theory of Statistics, Vol. 2: Inference and Relationship*, 443. New York, NY: Hafner Publishing Company.

[41] Aristophanes. *The Clouds*. (419 B.C.E. / 1938). In: *The Complete Greek Drama, vol. 2*, O'Neill, E., and Oates, W.J. (Eds.), 545. New York, NY: Random House.

[42] Meyer, J.J., and Veltman, F. (2007). Intelligent agents and common-sense reasoning. In: *Handbook of Modal Logic*, Studies in Logic and Practical Reasoning Series, Blackburn, P., van Benthem, J., and Wolter, F. (Eds.), 3(18):991-1029. Amsterdam and Boston: Elsevier.

[43] Trippas, D., Pennycook, G., Verde, M.F., and Handley, S.J. (2015). Better but still biased: Analytic cognitive style and belief bias. *Thinking & Reasoning,* 21(4):431-445.

[44] Marrin, S., and Torres, E. (2017). Improving how to think in intelligence analysis and medicine. *Intelligence and National Security,* 32(5):649-662.

[45] Ji, X., Niu, Y., and Shen, L. (2016). Robust satisficing decision making for unmanned aerial vehicle complex missions under severe uncertainty. *PLoS ONE,* 11(11):e0166448.

[46] Belnap, N.D., Jr., and Steel, T.B., Jr. (1976). *The Logic of Questions and Answers*. New Haven and London: Yale University Press.

[47] Safdari, R., Kadivar, M., Langarizadeh, M., Nejad, A.F., and Kermani, F. (2016). Developing a fuzzy expert system to predict the risk of neonatal death. *Acta Informatica Medica,* 24(1):34-37.

[48] Anderson, A.R., and Belnap, N.D., Jr. (1975). *Entailment: The Logic of Relevance and Necessity,* 1: 6-7. Princeton and London: Princeton University Press.

[49] Araya-Muñoz, D., Metzger, M.J., Wilson, A.M.W., and Carvajal, D. (2017). A spatial fuzzy logic approach to urban multi-hazard impact assessment in Concepción, Chile. *Science of the Total Environment*, 576:508-519.

[50] Fitting, M. (1995). Tableaus for many-valued logic. *Studia Logica,* 55:63-87.

[51] von Wright, G.H. (1957). Deontic logic. In: *Logical Studies*, 58-74. London, UK: Routledge and Kegan Paul.

[52] Kripke, S.A. (1963). Semantical analysis of modal logic I: Normal modal propositional calculi. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 9(5-6):67-96.

[53] Kripke, S.A. (1965). Semantical analysis of modal logic II: Non-normal modal propositional calculi. In: *Symposium on the Theory of Models*, 206-220. Addison, J.W., Henkin, L., and Tarski, A. (Eds.). Amsterdam, Netherlands: North-Holland.



Chapter 15 – WHEN TRADECRAFT WON'T WORK: DESCRIBING THE BOUNDS FOR ANALYTIC TECHNIQUE

James E. Kajdasz
United States Air Force
UNITED STATES

I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail.

Abraham Maslow

When a student of intelligence learns or becomes expert with a particular analytic technique, it is tempting to apply that technique to new intelligence questions that comes along. However, while the hammer is good for problems that are nails, they are less suited for screws, and no use at all on wing nuts. After learning ACH, I often tried to apply the technique in real-world problems. I sometimes became frustrated when the technique wasn't useful. In one instance an explosion had occurred in Indonesia. There was uncertainty as to the cause. Was it terrorism, or an accident? I immediately drew up my ACH matrix and began to enter the pieces of evidence as they were reported. No other explosions or gunfire were noted. The target did not have an obvious political value. The explosion seemed too big to be a gas leak. Terrorist groups had been threatening an attack. There was a history of terrorists setting off explosions of this sort. Additional hypotheses developed. Was it criminal activity? A government attack against a known terrorist location? The news networks brought on terrorism experts who each had at least a dozen new observations and opinions. My ACH matrix quickly became untenable. There was too much evidence of too little quality to make ACH a useful tool. ACH may not be much use at all (it lacks empirical support). Another reason for failure may be that we apply the technique in the wrong situation. It's reasonable to assume a particular Structured Analytic Technique (SAT) is most useful for a particular subset of problems. Hence, it may be helpful to develop a taxonomy that simply describes what families of techniques are useful for particular flavours of problems. To continue the analogy of tools: hammers are effective for pounding problems but not for twisting problems.

The goal of this chapter is to present a classification that assists in mapping analytic techniques to particular intelligence problems. Classification consists of grouping entities according to similarity on some set of criteria. As Bailey notes, the process is so central and ubiquitous to our everyday lives, that we are generally unaware of its existence. And yet, classification lays the foundation for conceptualization and language on a given topic. Without it, there can be no advanced conceptualization, reasoning, language, or data analysis. The secret to effective classification is to identify the fundamental characteristics to base a taxonomy or typology [1]. However, as Heuer and Pherson note, "there is no single right way to organize a taxonomy – only different ways that are more or less useful in achieving a specified goal" [2].

To be fair, most SATs come with a description for what set of circumstances they are best suited for. However, instructors and practitioners of intelligence analysis have a growing list of techniques: ACH, alternative futures, SWOT, Devil's Advocacy, utility matrices, Delphi technique, pre-mortem, etc. The ever-increasing list has grown extensive and is likely intimidating to new students, and burdensome to older analysts. As the number of techniques climbs, practitioners begin to specialize in a subset of techniques they have grown to appreciate and become expert at. The temptation grows to over-apply tools we know best. A taxonomy which maps techniques to problems serves to corral the unstructured list of techniques into a more digestible framework. It reduces complexity, highlights when a technique is unlikely to work, and suggests other techniques that are better suited.

There already exist typologies which classify both intelligence problems and analytic techniques. Treverton and Gabbard divide intelligence problems into puzzles and mysteries. Puzzles have a knowable answer that

can be determined once the correct information is obtained. A mystery is contingent on future influences, and cannot be determined at present with certainty [3]. Rittel and Webber made a formal distinction between tame problems, which scientific methods are largely successful at solving, and wicked problems that do not have a definitive solution, cannot be understood with methods of science, and are largely unique and one-shot affairs [4]. Friedrich Hayek introduced the concept of complex systems. According to Hayek, patterns can be discerned in complex environments, but not precise predictions as is possible with non-complex systems [5]. Most in the intelligence community have heard of Nassim Taleb's "Black Swan" problems, where improbable events can render prediction in complex environments an impossible task [6].

Taxonomies for SATs, or more generally problem-solving techniques, are less common, but still exist. Heuer and Pherson organize SATs in to eight categories according to analyst needs and the cognitive purpose they serve:

1) Decomposition and visualization;
2) Idea generation;
3) Scenarios and indicators;
4) Hypothesis generation and testing;
5) Assessment of cause-and-effect;
6) Challenge analysis;
7) Conflict management; and
8) Decision support [2].

Guilford makes a distinction between divergent and convergent thinking [7]. Divergent tasks (e.g., how many uses are there for a brick?) can have many acceptable answers. Imagination techniques, or challenge techniques, for example, are techniques designed to increase the available number of potential hypotheses. Hypothesis generation is a divergent thinking task, and analytic tradecraft techniques (e.g., alternative futures, brainstorming, red teaming) have been developed and categorized to assist with this kind of creative thinking. Convergent thinking tasks are those that have a single correct answer (e.g., 2 + 3 = 5). Selecting the single correct (or at least most likely) hypothesis from a list of possibilities is a convergent thinking task, and many analytic techniques have been developed to assist the analyst in choosing the single most likely option from a list of potential assessments (e.g., Team A / Team B, ACH, Bayesian Belief Networks, Prediction Markets).

There is nothing wrong with any of these taxonomies. All are useful in understanding the problem in some way. However, they don't specifically address the issue of matching tools of intelligence analysis to the intelligence questions based on the nature of the intelligence question. By "nature of the intelligence question" I am referring to the characteristics of the domain the intelligence question exists in (e.g., well understood problem with lots of data), rather than the goal of the technique (e.g., generate hypotheses), which is a more common way to classify techniques.

Snowden and Boone come closest to addressing the fit between nature of a problem and appropriate response [8]. In particular, they address leadership response. Managers must match their style to the complexity of the circumstances they face. Their Cynefin framework (named after the Welsh word for 'habitat') divides circumstances into four contexts: chaotic, complex, complicated, and simple. Managerial best practices differ depending on the domain. Chaotic circumstances, characterized by high turbulence and no clear cause-and-effect relationships, require a leader who takes immediate action to re-establish order, implementing what works as opposed to seeking right answers. When initial turbulence recedes, a chaotic environment may become a complex environment. Containing the characteristics of complex adaptive systems, cause-and-effect relationships are constantly changing. It may be impossible to find a "right"

answer, but with patience it may be possible to detect emerging patterns. If these patterns stabilize, a complex environment becomes a complicated one. A complicated context is characterized by stable cause-and-effect relationships not easily apparent, requiring experts to diagnose the situation. Appropriate manager behaviour includes creating panels of experts and listening to conflicting advice. When the cause-and-effect relationships are better understood, the situation may become a simple environment, with repeating patterns and consistent events. Leaders can apply fact-based management. But beware complacency. The old rules work until they don't work. When the old rules no longer apply, leaders may find themselves thrust back into the chaotic context. Thus, there is a cyclical aspect to this, although not all domains are capable of completing the cycle, as their progress is limited by both our understanding of the domain, as well as the characteristics of the domain itself. When trying to understand tribal allegiances, for instance, it may be that allegiances are constantly changing, and because of that the domain can never evolve beyond a complex context.

Just as Snowden and Boone use the context of a domain to suggest appropriate leadership behaviour, I believe the context can also suggest appropriate focus for the analyst. Prior to becoming aware of Snowden and Boone's work, I defined the context of the environment in terms of the value of theory and the value of data to answer questions in a particular domain. These two criteria, theory and data, form the horizontal and vertical axes of a framework denoting high/low theory application and high/low data application (Figure 15-1). I believe this categorization corresponds well to Snowden and Boone's Cynefin framework, where low theory / low data corresponds to the chaotic context, low theory / high data corresponds to the complex, high theory / high data to complicated, and lastly high theory / low data to the simple context.

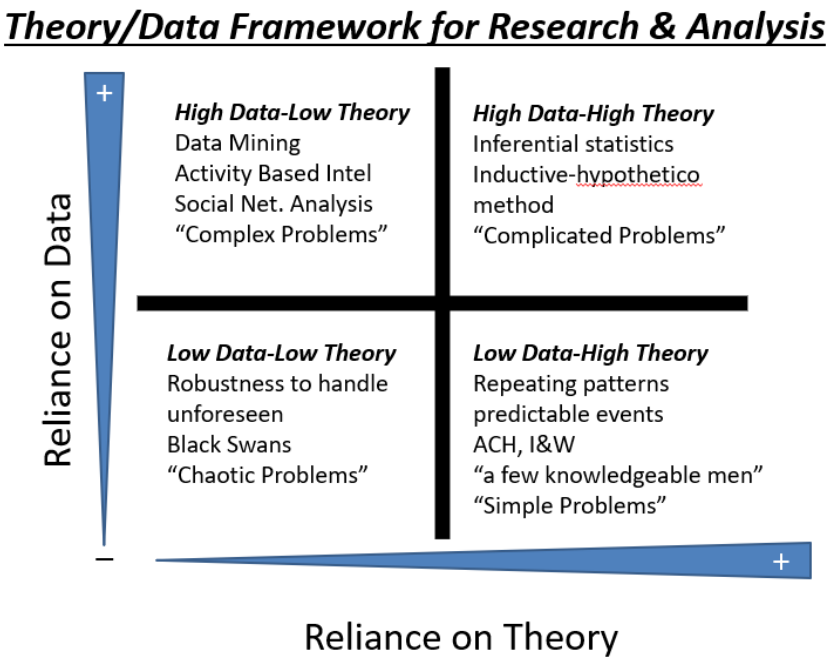


Figure 15-1: Theory/Data Framework for Research and Analysis.

Theory, in this chapter, is being used in the context of how the scientific community uses the word. Theory refers to some sort of organizing principle that explains the relationship between different variables. It is a system of ideas or a set of principles, often dealing with mechanisms or underlying reasons for behaviour, that help us organize and assimilate the observations that we observe. In this usage, a theory is more than "just an idea" [9]. A scientific theory "is grounded in actual data from prior research as well as numerous hypotheses that are consistent with the theory" [9]. Theories are general abstract statements that help to

explain the world around us. A theory is useful in contributing to understanding and knowledge, according to Pettigrew, in at least five important ways. Theories allow us to classify things. These "things" could be entities, processes, or causal relationships. Theories help us predict future events. They help us to understand what causes past events. Finally, they help guide research by providing focus on what questions should be explored more and what can be ignored [10]. A well-established theory is useful in explaining a great number of observable facts [9]. A theory is a general relationship that is relatively generalizable to a wide variety of circumstances. A theory that explains observations in one circumstance, to be useful, can also be used to explain other similar circumstances. Some intelligence questions may be unique one-of-a-kind events. Theory is less helpful in these cases. Other intelligence questions may be able to make better use of past similar events, or utilize general causal models to help predict what might occur. A model inevitably simplifies the real world by ignoring specific details or factors. However, models can have substantial value if they increase our understanding of the observations we make.

Data, as used in this chapter, is simply information and observations. Those observations might be quantitative or qualitative in nature. Some intelligence questions may have data available to draw upon. The explosion of social media and other online resources means that we are practically drowning in data for certain intelligence questions, and the challenge becomes getting and managing the useful data. Other intelligence questions may have relatively little data to draw upon because it is a recent, unfolding event or for some other reason has little information available.

The following examples illustrate how assumptions of theory and data can be used to understand different analytic approaches:

1) **Low Reliance on Theory, Low Reliance on Data:** Nasim Taleb's book, *The Black Swan*, fits well into this quadrant [6]. The Black Swan is a metaphor for a rare event that surprises people, and has a major impact. In hindsight, it is often rationalized as being inevitable or foreseeable. But, in actuality, such low-probability events are not predictable. When we make forecasts, we are making an estimate on what is most likely to occur. Black swan events are, by definition, improbable events. Things like the invention of the personal computer, the internet, the creation of nuclear weapons, the 11 September attacks, World War I… all of these things, Taleb argues, were highly improbable events that were essentially impossible to predict and it's futile to even try to predict these kinds of things.

   To Taleb's observation I'd like to add something I call the snowflake argument. The idea here is that a particular event can only be understood in the context of the local environment, and it's useless to try and apply what has happened elsewhere in time or place to the current set of circumstances. Each event is unique, like a snowflake, the likes of which has never been seen before, and never will be seen again. In this notion of knowledge, theory is not highly valued and data is not highly valued. If every event is unique, then nothing you know about the past is going to inform you about the present. And nothing you know about the present is going to inform you about the future. Theory concerns itself with the mechanisms that are generalizable from situation to situation. Theory predicts what is likely to happen. The Black Swan event, by definition, is something that's not likely to happen. Data might be useful only in retrospect, when determining the impact of a unique event, but it can't be used for prediction or to develop a more generalizable understanding of concepts.

   If you are an intelligence analyst, or an academic researcher, then this quadrant is a depressing quadrant to be in. There is little hope for increasing our understanding. There's little hope to advance knowledge, because each new event is a completely new set of circumstances. So, what analytic methods are good for answering analytic questions that are in this quadrant?

   If we find ourselves in Cynefin's chaotic context (which I suggest is analogous to the low theory / low data quadrant), Snowden and Boone suggest a limited opportunity for thoughtful analysis. Many quick decisions will need to be made with little time for consideration. Snowden and Boone also note chaos can be an opportunity to implement innovative changes in a system. Analysts

might highlight such opportunities to decision makers. Ultimately the goal is abating the immediate crisis so one can move from a chaotic context to a complex context. This requires satisficing. "Doing what works" quickly is better than "doing what's right" slowly.

Taleb recommends the only response to a Black Swan event is to develop robust systems that are capable of rebounding to a shock. If an analyst is asked to prepare an analysis on the next Black Swan event, the appropriate response might be to look inward, and evaluate the critical points of failure in one's own organization. Rather than trying to predict or understand the nature of a future shock event, (e.g., will the price of oil spike next year?), the analyst instead suggests 'what if' scenarios that the organization would need to respond to (what if oil suddenly spiked?) and the organization can reflect on how they would be affected. If significant weaknesses are uncovered, systems can be modified to make them more robust to future unpredicted shocks.

2) **Low Reliance on Theory, High Reliance on Data:** This quadrant is analogous to Snowden and Boone's complex context. Theory may be lacking because the domain is too new to be well understood. Or characteristics of the environment mean cause-and-effect relationships are constantly changing, as is often the case in complex adaptive systems, making stable, generalizable theories rare. Some domains may never progress beyond complexity due to the characteristics of the environment. In such cases the search is for stable patterns that will likely have temporary usefulness.

What methods might be useful in this domain? There are data-centric, non-theoretical approaches. One is Activity-Based Intelligence (ABI). Instead of focusing collection on specific objects over a small period of time, this intelligence methodology focuses on collecting data on activities and transactions over long periods of time and large areas [11]. If you build a picture of what "normal" activity looks like, you can spot when something abnormal occurs, and cue a human analyst to take a look at that particular event more closely. Machine learning and data mining can be a data-centric approach as well, where one goes into the maelstrom of data looking for relationships without a theory to guide one's search.

There is opportunity for a data-centric approach to go wrong. A caution to the data-centric approach is presented by the story of Paul the Octopus. Paul the Octopus lived in an aquarium in Germany and during soccer matches in 2008 and 2010 Paul made predictions of which team would win by eating tasty octopus treats out of one of two possible boxes. He was correct 11 out of 13 times for an impressive success rate of 85%! The oracle Octopus even beat Goldman Sachs analysts who used data to make their own predictions [12]. However, if Paul the Octopus made a new prediction for a future game, I personally would still not place a bet based on his input. Paul may have a good record of picking the winner, but there's no theoretical reason why Paul's choice should be any better than a coin flip. Paul the Octopus highlights one of the dangers of a purely data-centric approach. If you look at a large number of variables, some of them will correlate with each other just by chance. If you look at enough octopi, some of them will pick winners.

Taleb [13] highlights this danger: If you take a large number of completely random variables, some of those random variables will correlate just by chance. A set of 200 variables with 1,000 observations each will, for example, almost certainly have at least a few significant correlations. These chance correlations, or spurious correlations, are one of the dangers of using a data-centric approach that does not take into account theory. As you increase the number of variables (even random variables) the more chance correlations you will find.

If patterns can be found and these patterns are relatively stable, they can be used to create theories. These theories can be applied to understand other times, places, and situations. If this is the case, we can progress to a high theory / high data environment.

3) **High Reliance on Theory, High Reliance on Data:** If both theory and data can be leveraged, we are in a quadrant analogous to Snowden and Boone's complicated context. Cause-and-effect relationships are relatively stable, but not obvious. Experts are required to uncover the connections. This is close to the traditional scientific method and we have developed excellent methods for this context.

But how might this apply to problems in intelligence? As an illustrative example, let's say you deploy to Afghanistan. While there, some of the local roads get paved. You notice that the insurgency gets worse after the roads are paved. You make that observation and come to a belief that paving roads makes insurgencies worse. You took your Afghanistan experience (data), and you believe you learned something about all insurgencies in general (theory). We can take theory and attempt to validate it with data. Let's take the effect of paved roads on insurgency. We could test that theory by looking at other insurgencies. Our theory makes a prediction that the insurgency will be worse in those areas that have paved roads.

This process is the hypothetico-deductive model, familiar to most scientists. We start out with an observation such as our deployment to Afghanistan. We use that experience (that data) to generate a theory that paving roads makes insurgencies worse. This is inductive reasoning where we take one observation and generalize the observation to the world. We take our theory that paving roads worsens an insurgency, and we make a prediction about what the data will show in other insurgencies. If what I observed in Afghanistan is true of all insurgencies, I can make a prediction that I will see a similar relationship in other insurgencies like Egypt or Syria or Libya (deductive reasoning). The insurgency violence will be worse in those areas with paved roads. Ideally, I could do an experiment by randomly choosing some insurgency areas to be paved, and others not to be, and observe the effects at a later time.

This is a continuous cycle. Scientists go from data to theory to data to theory… And as the cycle continues, theories are refined, and they become better at explaining the world around us. We may find that our paved-roads theory needs to be modified. We might refine our theory by saying paved roads make the insurgency violence worse, but only in areas where it is not possible to travel across open ground easily. The revised theory, if it is a good revision, will be able to explain more data. The more observations we're able to explain with our theory the more useful that theory is.

This mechanism allows human knowledge to advance. Theories usually outlive their originators, and so it's through the revision of better theories that a discipline evolves and becomes better at what it does. The next generation of scholars and analysts can stand on our shoulders and advance the discipline. Otherwise, if we don't advance theory, if we don't stand on the shoulders of those that have come before us, the nature of wisdom stays one-generational. The future generation of analysts can only hope to be as good as the current generation of analysts.

An example of a data-strong / theory-strong approach in intelligence is Richard Cincotta's application of youth bulge theory. Youth bulge theory notes that the 16-to-30-year age range is associated with risk-taking, especially among males. Youth bulges occurring in this age group in developing countries are associated with higher unemployment and, as a result, a heightened risk of violence and political instability. So, youth bulge theory argues that an excess of young adult males predictably leads to social unrest, war and terrorism.

Cincotta wanted to add to this theory. He thought that youthful demographics would be more likely to have authoritarian regimes. After all, if you are a shop keeper in a country with large numbers of young, unemployed, risk-taking males, you may not really mind that the police have some extraordinary powers. But as a demography matures, these concerns fade and issues of freedom and civil rights become more important. What's interesting then is Cincotta made projections of future

populations, and identified which countries are most likely to go through this transitional phase from having an authoritarian style government, to having a form of government more like a liberal democracy [14]. He made this prediction based only on the demographic profiles. The first region that promised a shift to liberal democracy was a cluster along Africa's Mediterranean coast: Morocco, Algeria, Tunisia, Libya, and Egypt, none of which had ever experienced democracy in the recent past. We saw some of this transition potentially start during the Arab Spring. The other area is in South America: Ecuador, Columbia, and Venezuela, each of which attained liberal democracy demographically 'early' but was unable to sustain it. For the policymaker, these areas represent opportunities where US efforts stand a good chance of bringing more democratic forms of government. It also suggests that countries like Afghanistan are just too young demographically to see a democratic form of government anytime soon.

As theory becomes better at explaining the world around us, we may find we don't need to collect as much data as we used to. Theory can direct us to a few key pieces of information that will tell us most of what we want to know. If we can reach this stage, we transition to the high theory / low data quadrant, analogous to Snowden and Boone's simple context.

4) **High Reliance on Theory, Low Reliance on Data:** Mark Lowenthal [15] argues that the intelligence community has focused too much on text-based data manipulation and not enough on critical thinking, experience, and wisdom. Lowenthal doesn't claim that data are unimportant. However, as the title of his article ("A few knowledgeable men") suggests, you don't need huge numbers of people to look at all the data. Rather, you just need a smaller number of men and women who are knowledgeable about the background issues for a particular intelligence question. What you need is a few people who understand the underlying motivations, culture and history. If you just have a few knowledgeable men and women who understand theory, then you don't need to look at every piece of data. You understand the direction the wave of data is traveling even though you aren't examining every piece of data. Data is deemphasized in this approach.

   I'd like to highlight a risk to this approach, however. We are good at creating a quick story to retroactively explain past results. An example I use in class is the earlier-mentioned theory about paving roads in Afghanistan. Sometimes in class I describe how we are actually helping the insurgency by paving roads. This makes a lot of sense to the students. Allied forces use a lot of air support and air transportation. The insurgents have to use ground transport. So, the roads are actually helping them more than they are helping us. Insurgents can now more quickly and more efficiently move supplies and forces. In addition, allied forces are now using the roads a lot more too, and not relying on air transport, so it presents a concentrated target that the insurgency can take advantage of.

   Sometimes I present the exact opposite finding in class. Paving roads in Afghanistan is hurting the insurgents. People are also able to easily explain these results as well. Well of course it hurts the insurgency! By building roads, you are building up the transportation infrastructure of the country. You're facilitating trade of goods. You're facilitating the growth of business which is going to help grow the economy of Afghanistan and give young Afghans options that they would not have had otherwise. Now there are things you can do with your life other than join the insurgency. The point here is that an idea that is not supported with data is just an idea. And many people can have conflicting ideas. The data should help you decide which opinion is correct.

   If you have a well-established theory, and it is well supported, it can be used to make future predictions, or bring understanding. This is especially useful when data is not available or time is short. However, Snowden and Boone have a strong warning to organizations operating in the simple domain. The old rules work until they don't. They recommend challenging orthodoxy and avoiding complacency.

So, which of these views of knowledge is the correct one for intelligence analysis and for academic research? All of them have a place. Looking at the "snowflake" quadrant for a moment, we have to acknowledge that every event does have some aspects that are truly unique. For some events, truly nothing like them has ever been seen before or will ever be seen again. Amy Zegart gives the example of a nuclear Iran [16]. What would a nuclear Iran look like? What precedence should we look to in history for estimates on how a nuclear Iran might act? Only nine countries have nuclear weapons. Five got the bomb a very long time ago when the world climate was very different. Two of the more recent nations to acquire the bomb also seem like poor models: there's North Korea, which hardly seems a generalizable model for anything, and South Africa, the only country that developed and then voluntarily dismantled its nuclear arsenal. Intelligence analysts don't have a rich historical store of comparable cases to help assess future outcomes for Iran.

Data mining, even without theory, can play a vital role in intelligence in the area of indications and warnings, for instance. If we were to observe large amounts of activity over a very wide geographic area, and then create a picture of what normal activity looks like, then we can identify and highlight non-normal activity when it occurs. That could be a significant tool for warning.

A few knowledgeable men or women who know about theory can be very helpful when data is sparse or when time is short. The whole purpose of theory is to be able to give us the ability to understand the relationships in a broad sense without having to look at reams and reams of data. If our theories have been well validated, then a few knowledgeable men and women are very useful in understanding the world around us.

There is a rich history of applying the scientific method to both data and theory. We have this wonderful potential of building theory over time such that future generations of intelligence analysts can do a better job than we are doing today. That's very powerful, but it also relies on some assumptions. It assumes that some relationships are going to be generalizable from situation to situation. That is to say, it assumes that there is the possibility of developing theory. It also assumes we will have data that we are capable of observing. Sometimes historians debate the circumstances surrounding particular events for decades. Sometimes data is not readily observable.

5) **How systems change over time:** The power of the theory-data framework comes into play when systemic change to the nature of a problem over time is considered. Let's assume the environment we are operating in has undergone a "Black Swan" shock of some sort. Examples of this kind of shock might be Iraq invading Kuwait, or 9/11. We find ourselves in the chaotic quadrant of low data / low theory approaches. This is not the time for high data approaches such as data mining, or inferential statistics. The commander in this instance must act quickly with little information to stabilize the situation, stem further bleeding, and attempt to bring order. For example, after the invasion of Kuwait, US fighters were sent to sit on the tarmacs of Saudi Arabia to dissuade Iraq from continuing their invasion into Saudi Arabia. After 9/11, all commercial air traffic was grounded.

   Once the environment stabilizes somewhat, we may find we have progressed to the high data / low theory quadrant of complexity. Different actors are manoeuvring for power and changes can have large, non-linear effects on the overall system. In this instance, the analyst is best served by high data approaches such as Social Network Analysis, which is useful in identifying who the major players are in the current system. Data mining helps us understand the new system we find ourselves in. Theory is hard to apply in the frequently changing dynamics of the current system. Arguably, many intelligence questions remain in this realm.

   Perhaps, over time, the environment proceeds to a state of more stable institutions and power bases, rendering the environment a bit more predictable and allowing theory to make predictions over a

wider variety of observations. We find ourselves operating in the complicated realm, where high data / high theory approaches will be helpful. Linear regressions can be used to make consistent prediction of how much wheat Country A will produce, based on the institution's stated goals and a few environmental factors. Number of tanks produced can be estimated, based on imagery of production facilities and orders of raw materials. Experts are helpful in this realm, and high data / high theory techniques that help to aggregate expert opinion, such as Bayesian Belief Networks or Prediction Markets, may be helpful here.

Finally, certain problems may allow us to proceed to the last high theory / low data approach. In this quadrant, the problem is well understood, and we can rely on what has happened in the past to make good predictions about the future. Key pieces of data can be accessed, rather than attempting to look at all data. For instance, Country A has conducted many weapon tests in the past. We know activity at a few key buildings precipitates the next test, and we can reference those few buildings to predict future tests. This is the realm of "a few knowledgeable men" who have the background to know what to look for. They are able to make accurate predictions by looking at perhaps just a few key diagnostic pieces of information. ACH might work better here, since we can pre-populate a matrix with the key evidence we know to look for. Stable and predictive Indications and Warnings matrices can be developed.

In Snowden and Boone's Cynefin Framework, they warn about the simple/chaotic boundary. When operating in the simple domain it is easy to let complacency set in and the environment may change without our notice. In our example of predicting weapon testing, imagine new R&D buildings have been constructed that we are not aware of, and weapons testing has transitioned to those buildings. The old theories that worked in the simple domain no longer apply. We are launched into the chaotic realm to start over.

I'm not certain which SATs will work best in a particular quadrant. I do have some intuitions. If the domain of your intelligence question has become chaotic, I suspect it's not the right time for a linear regression. Save classic statistical analysis for domains in the complicated context. If you are in the complex domain, strategies that locate temporary patterns to exploit will be more helpful than attempting to validate theories. If your intelligence problem is in the simple domain, this might be a warning sign that complacency has crept in, and you should take measures to look at new data. My hope is that presentation of this framework provides vocabulary to discuss appropriate analytic focus for a given intelligence question.

15.1 REFERENCES

[1] Bailey, K.D. (1994). *Typologies and Taxonomies: An Introduction to Classification Techniques.* Thousand Oaks, CA: Sage Publications.

[2] Heuer, R.J., and Pherson, R.H. (2011). *Structured Analytic Techniques for Intelligence Analysis.* Washington, DC: CQ Press.

[3] Treverton, G.F., and Gabbard, C.B. (2008). *Assessing the Tradecraft of Intelligence Analysis*. Technical Report Series. Santa Monica, CA: RAND Corp. https://www.rand.org/pubs/technical_reports/TR293.html.

[4] Rittel, H.W.J., and Webber, M.M. (1973). Dilemmas in a general theory of planning. *Policy Sciences,* 4:155-169.

[5] Hayek, F.A. (1974, December 11). *Prize lecture: The pretense of knowledge*. Retrieved 11 March 2016 from http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1974/hayek-lecture.html.

[6] Taleb, N.N. (2010). *The Black Swan: The Impact of the Highly Improbable* (2nd ed., trade pbk. ed.). New York, NY: Random House Trade Paperbacks.

[7] Guilford, J.P. (1967). Creativity: Yesterday, today and tomorrow. *Journal of Creative Behavior,* 1(1):3-14.

[8] Snowden, D.J., and Boone, M.E. (2007). A leader's framework for decision making. *Harvard Business Review*, November: 69-76.

[9] Cozby, P.C., and Bates, S. (2012). Where to start. In: *Methods in Behavioral Research*, (11th ed.), 18-39. New York, NY: McGraw-Hill.

[10] Pettigrew, T.F. (1996). Thinking theoretically. In: *How to Think Like a Social Scientist*. Santa Cruz, CA: Harper Collins Publishers.

[11] Long, L. (2013). Activity based intelligence: Understanding the unknown. *Intelligencer: Journal of U.S. Intelligence Studies,* 20(2):7-15.

[12] Chaturvedi, N. (2014, June 27). Octopus beats "vampire squid" as Goldman falls in World Cup. Retrieved from http://blogs.wsj.com/moneybeat/2014/06/27/goldman-sachs-loses-in-vampire-squid-vs-octopus/ (14 March 2016).

[13] Taleb, N.N. (2013, February 8). Beware the big errors of "big data." Retrieved 8 April 2013 from http://www.wired.com/opinion/2013/02/big-data-means-big-errors-people/.

[14] Cincotta, R. (2008). How democracies grow up. *Foreign Policy*, March/April:80-82.

[15] Lowenthal, M.M. (2012). A few knowledgeable men. *The American Interest*, Spring:87-89.

[16] Zegart, A. (2013, April 13). No one saw this coming: How we've been surprised by our growing ability to predict things. Retrieved from https://foreignpolicy.com/2013/03/13/no-one-saw-this-coming/ (20 October 2018).

Part IV: COMMUNICATING UNCERTAINTY IN INTELLIGENCE PRODUCTION



Chapter 16 – ISSUES OF UNCERTAINTY IN NATURAL LANGUAGE COMMUNICATIONS

Kellyn Rein
Fraunhofer FKIE
GERMANY

It appears, from all this, that our eyes are uncertain. Two persons look at the same clock and there is a difference of two or three minutes in their reading of the time. One has a tendency to put back the hands, the other to advance them. Let us not too confidently try to play the part of the third person who wishes to set the first two aright; it may well happen that we are mistaken in turn. Besides, in our daily life, we have less need of certainty than of a certain approximation to certainty.

Remy de Gourmont, Philosophic Nights in Paris [1]

16.1 INTRODUCTION

Information plays a central role in surviving and coping with today's world. Effective decision making depends on the quality, completeness and trustworthiness of the information which decision makers have at their disposal. Actionable intelligence for situational understanding is garnered both from human sources, in the form of text or speech, and from devices such as radar, acoustic arrays, ground sensors, video, etc., which report data on physical phenomenon in either digital or analog form. Both human and non-human sources may pass on data which is not fully accurate or trustworthy; however, there are some significant differences in the causes of the inaccuracies; understanding and analyzing these differences appropriately can improve the quality of the intelligence received.

A sensing device such as a video camera, a radar dish or a ground sensor is a neutral observer of events taking place within its scope. A video camera records light waves (and, if so equipped, also audio), a radar dish gathers data about movements with the range of its sweep, and a ground sensor documents vibrations detected near its location. Data from devices is always historical, that is, the physical events – motion, temperature, light, etc. – recorded by the sensor have actually taken place (fusion algorithms based upon data received from the sensor may project future states, but these are independent of the sensor itself). Sensor data is also neutral and impartial; the sensor has no vested interest in the meaning of its recordings. Furthermore, the type of data delivered by a device is always the same; the device records only certain types of physical phenomena. A thermometer does not measure motion, a motion detector does not measure temperature, an acoustic sensor does not record light waves.

In contrast, humans are multi-purpose sensors. We see, hear, feel, taste and smell, and we communicate what we have sensed using natural language. Our capabilities in each of these observational modes vary from person to person based upon a number of factors including physical condition (e.g., excellent or poor eyesight?), background knowledge in the phenomenon observed (e.g., trained expert or layperson?) as well as expectations or assumptions about the phenomenon which is being observed (is this normal or out of the ordinary?). In other words, humans often, intentionally or unintentionally, pre-process, interpret, speculate on or self-filter their observations thereby modifying the observation before passing it on. Furthermore, humans relay information about events which they have not personally experienced or observed, for example, in the form of hearsay or in discussion of future events which have not yet taken place. In addition, humans pass on opinion, speculation, assumption and inferences. But one of the most significant differences between a device and a human is that, whereas the device may provide bad data due to malfunction or environmental factors, humans can – and do – lie, distort or otherwise misrepresent information ("fake news"). Human beings, as sources, are therefore problematic on several levels, which we will discuss in further depth in a later section in this chapter.

Regardless of the type of source from which the information comes, it is vitally important for decision makers to have a realistic idea of how good that information is. Ideally, we would only use information which is complete and which has been definitively confirmed as factual. However, the reality is that this is seldom the case, and that we often need to make decisions based on incomplete and uncertain information. While using uncertain, incomplete, misleading or incorrect information as the basis for action can be a recipe for disaster, there are many times when incomplete and uncertain information is all that is available to decision makers. Realistically assessed, even partial and uncertain information can be used to great effect. For example, in 2011 when President Barack Obama gave the order for an assassination attempt on Osama Bin Laden at his hideout in Pakistan, the President himself rated the odds of actually finding Bin Laden there as "50-50" [2].

Device-derived data may be uncertain or unreliable for a variety of reasons. For example, sensors may be affected by environmental conditions such as heat, humidity or light conditions thereby producing unreliable data. Devices may also malfunction and fail. However, devices may be tested and calibrated under various conditions, giving decision makers important information about the overall reliability of a given source under different physical situations. Also, the algorithms which operate upon device-derived data have been a research focus for several decades now and are quite mature and well understood. While there is still much work to be done, on many levels, this work focuses on refinement, improvement and tweaking of existing technologies. (There are many fine works which provide excellent coverage of the advances in this field.)

Human-derived data, in contrast, remains problematic, in part because of the factors mentioned previously, but also because the information humans convey is delivered in natural language. Natural languages are flexible enough to deal with almost every aspect of the human experience and thus are powerful communication tools. At the same time, the flexible power of natural language information often makes it inherently uncertain, in part because natural language utterances are often ambiguous, vague, open to (mis)interpretation, or even incorrect. However, because uncertain information can play a vital role in intelligence analysis, in particular in the prediction of events which may happen in the future, we need to accept this uncertainty and find strategies to deal with it appropriately.

Thus, it is important to identify and understand how natural language information content itself is less than certain. Sometimes the uncertainty has to do with the content itself in the form of ambiguity ("I saw her duck"), vagueness ("down the road") or imprecision ("some", "tall", "many"), which may be context-dependent. Often uncertainty is conveyed by the speaker in the form of lexical forms that express the writer's stance toward the truth of the proposition in the sentence ("unlikely", "possibly"), indicate the origin of non-observed information ("people say", "I assume"), and other constructs such as modal verbs or future tense ("might", "will be"). For intelligence purposes, it is important to differentiate between uncertainty within the proposition and uncertainty about the proposition.

Our focus in this chapter is to look how and in which ways natural language is uncertain, as mentioned previously, and look at how these affect the quality of the reporting. Finally, because making sense of the huge volume of natural language information being generated on a daily basis requires some automatic pre-processing (e.g., text analytics) to locate potentially useful information, we will present a methodology identifying, evaluating, and weighting the evidentiality of textual information, with particular emphasis on lexical markers which the source used to convey the origin of the information being passed on, as well as their assessment of the quality of that information.

With the huge volume of natural language information being produced on a daily basis there is an ever-increasing reliance on computers for text analysis to discover actionable information buried in the flood of words. Text analytics utilize algorithms to locate identify certain patterns embedded in text which may identify objects or individuals, events, relationships and other useful information. The weakness of these algorithms is that they seldom, if ever, take into account that some of those patterns may be couched in language that indicates the discovered events or relationships may be questionable or even false. In other

words, for all intents and purposes all information extracted using text analytics is treated as "fact," even though there may be clear evidence buried in the surrounding that they are not.

Thus, in order to be truly actionable information, a parallel process to text analytics should be carried out, in which evidence of uncertainty expressed in lexical forms, when found, is analyzed and utilized to assign an initial credibility rating.

In the following sections we will look at uncertainty in text, with the major focus on those forms of uncertainty which may help us evaluate the credibility of the information we receive.

16.2 OVERVIEW OF UNCERTAINTY IN TEXT

As already briefly touched upon, the information ("signal") delivered by a human source is in the form of spoken or written text. Some statements are precise, accurate historical accounts of actual, directly observed events. However, a great deal of human communication is neither precise nor accurate nor historical and therefore may be uncertain in one or more aspects.

Consider the following sentence:

1) I think someone said there were some animals in the road.

There are several ways in which this sentence is uncertain. The statement begins with I think, which expresses belief or opinion, not knowledge; the speaker is letting the listener know that there is some doubt about the veracity of the rest of the statement. Someone said indicates that the assertion is hearsay, and therefore second-hand information, which may or may not have been correctly understood by the speaker and therefore not an accurate reflection of what the original source had, in fact, said.

Within the assertion itself, there are also elements of uncertainty (there were some animals in the road). For example, we do not know how many some is, nor do we know what type of animals there were, nor even which road is being referred to (although this may have been determinable from the text surrounding this statement, i.e., from context knowledge).

Although all of these reflect uncertainty, the elements mentioned above differ as to the uncertainty that they project within the statement. In fact, there are two basic categories of detectable uncertainty which appear at the sentence-level in written text or speech:

- Uncertainty within the content, including:
  - Imprecision;
  - Vagueness; and
  - Ambiguity and polysemy.
- Uncertainty about the content, including:
  - Modal verbs;
  - Modal adverbs (including "words of estimative probability");
  - Hearsay markers;
  - "Mindsay" markers indicating belief, inference, assumption, etc.; and
  - Passive voice.

While "hearsay" is a familiar term, "mindsay" may be new to many readers. This term appears in Ref. [3] and is used to describe information which is not a result of direct observation nor of passing on second-hand information, but which is the product of someone's mind, e.g., inference or belief.

The first category – uncertainty within the content – is important to applications such as information fusion, in which assertions containing imprecise or vague descriptions may be analyzed to determine whether these assertions refer to the same or to different objects. For example, suppose we are searching for an individual described by police as 5 feet 9 inches (175 cm) tall with light brown hair – when we receive a report about a tall, blond man, how likely is it that the report is about the suspect? At what height is a person "tall"? When is hair light enough to be considered blond rather than light brown? The potential ranges of values for vague or imprecise terms are often dependent upon a specific domain. We will discuss these in more depth in the following section.

Uncertainty about the content is a type of non-content meta-information delivered by the speaker to the listener which provides important clues about both the original source of the information,; for example, if the speaker is describing transmitting second- or third-hand information, or if the assertion is the result of opinion, belief or a logical process.

In Ref. [4] we classified the uncertainty within the content as "inexactness", and the uncertainty about the content as "evidentiality" as shown in Figure 16-1. The latter type of uncertainty provides clues about how strongly or weakly the content of the sentence may be considered as "evidence" of the state or event described.

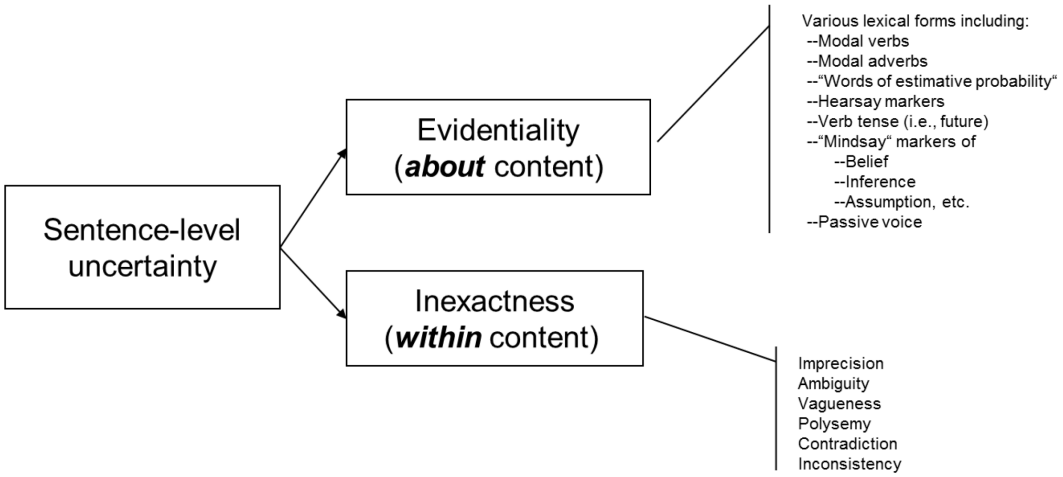


Figure 16-1: Sentence-Level Uncertainty in Natural Language Communications.

In the following subsections we will examine both types of uncertainty in more depth. We begin with inexactness, which is important in helping to determine whether or not two or more reports refer to the same object or event – in other words, for corroboration. After that we will discuss evidentiality, which plays a role in the credibility of reported information.

16.2.1 Uncertainty Within the Content

When humans communicate, we do more than convey basic facts. We express thoughts, hopes and wishes, we speculate about the future, we pass on information that others have communicated to us; we tell lies and half-truths to elicit cooperation, to be accepted as part of a group, to win approval from others or to evade censure. Even when we are in fact passing on concrete information, we don't necessarily deliver it in clear, concise and

precise wording. We may use words that have multiple meanings or formulate our sentences so that they are ambiguous. There are a variety of ways in which the informational content ("signal") can be uncertain which we will examine in this subsection.

16.2.1.1 Imprecision and Vagueness

Human communication is often formulated in ways that obscure, however unintentionally, details that may be useful in the information fusion process.

Let's return to the informational assertion of the example used above:

2) There were some animals in the road.

Some is an imprecise number. The reader might possibly proffer judgements on the range of numbers represented by *some*: it definitely means *more than one*, but the upper bound is unclear. Suppose the speaker would have used *a couple* as a descriptor; in general, this refers to a small number, which, while still greater than one, is most likely a very small number, say, two or three. Use of *a bunch* indicates more than *a couple*, whereas *many* would have been preferred if there were a large quantity. *Several* is usually understood to be more than *a couple* – perhaps five or six – but generally would not be considered to be *many*. However, *some* simply implies *multiple* without any further hint as to exactly how many, so we can only guess.

Additionally, it is not clear what sort of animals these are: chickens? dogs? cows? camels? Quite possibly it was a mixture of different types (a sheepdog and a dozen sheep, for example). One type of animal would most likely not be intended here: a human. In that case, some people would have been used in place of some animals. However, the statement does not necessarily exclude the presence of a human: for example, when a herd of sheep are in the road, the shepherd is often somewhere in the vicinity as well, but the animals, not the shepherd, would likely be considered as some sort of anomaly worth mentioning. Similarly, a high level of precise detail may also be inaccurate: even if we were told that there were six brown Jersey cows in the road, it may well be that the observer neglected to let us know that there were two border collie dogs as well, or overlooked that one brown cow was, in fact, a Hereford and not a Jersey.

(It should also be noted that presence or absence of precise details may often be considered by experienced analysts as indicators of fabrications or lies.)

Correlating information about a shepherd and his dogs moving his cattle to new pasturage with some animals in the road requires understanding of a variety of things, including that *dogs* and *cattle* are animals, and that *some* means multiple.

Complicating things further, many vague or imprecise formulations may be context or domain dependent. For example, large is an adjective related to size of an object, but its exact (quantifiable) meaning is extremely domain dependent. There are many orders of magnitude difference in the numerical values indicated by large between a large city, a large ship, a large dog and a large molecule.

Furthermore, even within the same domain, there may be variations due to other factors such as context information. For example, the phrase a lot of people will generate a different numerical range depending on expectation or physical factors such as facility size. If a smallish meeting room is filled to standing room only, it will be reported that the 50 persons attending the event were a lot of people. However, those same 50 persons would not be classified as a lot of people if they are the only occupants of a 400-seat auditorium. Likewise, 50 people in a 30,000-seat sport stadium would generate a comment more like nobody was there. Therefore, the decision about the numerical range represented relies on what we know about the location. Gross *et al.* [5] have discussed such problems and their resolution at some length.

16.2.1.2 Ambiguity and Polysemy

Statements may be ambiguous, that is, they may be open to more than one interpretation or have more than one possible meaning:

3) Students hate annoying professors.

4) Sally gave Mary her book.

In statement 3) it is unclear whether the students strongly dislike professors who irritate them, or whether students try to avoid making their professors angry, perhaps in the hope of receiving a better grade in the course. In statement 4) we do not know whether Sally gave Mary her own (Sally's) book, or whether Sally was returning Mary's book to her – or even if there may yet be another female involved. For example, it is possible that in a preceding sentence, the writer informs us that Jane is a well-known author and a friend of Sally's, and thus 4) tell us that Sally presented Mary with a copy of Jane's latest hit novel.

Another example of ambiguity is shown in 5):

5) I saw her duck.

There are two possible interpretations of statement 5); the first is that the female person referred to has lowered her head to avoid, say, a low-hanging branch, the second is that she keeps a waterfowl as a pet. This ambiguity stems from what is called polysemy, the fact that the same word (label) can refer to two or more separate concepts. Other examples of polysemy are words such as bank which could be a financial institution, the side of a river, or a motion executed by a flying aircraft, to name just a few of the many meanings. Determining the intended meaning can generally be achieved by analysis of the surrounding text.

Regardless of the lack of detail in all of the examples above, there is nothing in any of these sentences to make us believe that the sentences are not true. However, very often sentences contain clues which the writer uses to signal to us there may be reason to doubt the veracity of the content contained in the sentence. These we will discuss in the following section.

16.2.1.3 Which Language?

Last, but not least, one of the most obvious problems is quite straightforward, indeed almost trivial: there are a multitude of spoken and written natural languages. According to the Linguistic Society of America, there are nearly 7000 distinct natural languages, of which some 230 are spoken in Europe [6].

But the problem is not just limited to diverse languages. A language such as English has a variety of regional variants with noticeable differences. The Irish playwright George Bernard Shaw is credited with commenting that England and America are two countries separated by a language in common. This is not simply a matter of pronunciation or even spelling – it is also a matter of vocabulary. The *biscuit* of a Brit is an American's *cookie*, and an American's *biscuit* is more akin to an unsweetened British *scone*. Here we have an instance of the same word being associated with two different concepts based upon the variant of English being used – an example of the polysemy discussed in Subsection 16.2.1.2. There are also some lexical differences: if you ask an American about a *lorry* you will get a blank look – for her, the object referred to is a *truck* and the word *lorry* does not exist in the American variant. Here we have two separate words for the same object – in essence, synonyms, but only in a cross-variant sense. Therefore, it is essential to know which version of English is being used.

In addition to regional language variants, there may be examples of polysemy which are domain-specific, that is, they may be very unique to certain subgroups of native speakers of that language, but not to all speakers. For example, "POV" within the US military is used as an abbreviation for "privately owned vehicle" (i.e., a soldier's own car), but within the fiction community, writers often use

"POV" to mean "point of view." Therefore, the meaning attached to the acronym will vary according to the context in which it appears. For outsiders, these need to first be deciphered or disambiguated.

Another issue, particularly with an internationally widely used language such as English has to do with irregularities produced by non-native speakers which can also cause misunderstanding and communication problems. This is often caused by what is known as "false friends" or words appear to be identical in both languages, but have quite different meanings. For example, native German speakers often misuse the English word *actual* (meaning *real* or *existing*) when they mean *current* (as in *current news*) because the German word *aktuell* has the latter meaning. Another variation is that non-native speakers borrow terms verbatim from another language but assign them a meaning not existing in the original language: an American uses *old-timer* to refer to an old person, whereas a German using *oldtimer* means a classic auto.

Sometimes the issue is that languages do not map concepts one-to-one. In American English there are three words – *pumpkin*, *squash* and *gourd* – for a related family of fruits which map to only two words *Kürbis* and *Zierkürbis* in German. While *Zierkürbis* maps one-to-one onto *gourd*, *Kürbis* is used for both *pumpkin* and *squash*, leading to confusion for native speakers of English.

Yet another phenomenon is the invention of words which do not exist in the native version of the language. Ask a native speaker of any English variant what a *pullunder* is and they will be puzzled, whereas a German speaker will describe to you a sleeveless sweater known as a *sweater vest* in the US and Canada, and as a *slip-over* in England.

And finally, there is the issue of "code switching" [7] in which multilingual speakers use different languages within a single communication. It is not uncommon that a natural language acquires words or phrases from other languages, which then become standardized vocabulary for the acquiring language. Examples of this in English are *gestalt* and *angst* (from the German *Gestalt* and *Angst*), which are now fully integrated into the lexicon of the acquiring language, and can no longer be considered "foreign."

In contrast, "code switching" means that the speaker shifts from one language to another within a sentence (i.e., a word or phrase) or between sentences in a single communication. For example, it is not unusual in expatriate communities for speakers to insert words or phrases of a language from the host country into their mother tongue communications, particularly with other expatriates who likewise understand the foreign words. This can be problematic for automatic text analysis, which relies on vocabulary and grammatical structures of a single language and can cause confusion or uncertainty as to the meaning of the sentence, through uncertainty of individual words or phrases.

Clearly, there are a number of significant challenges to understanding the meaning of the information received from human sources. But regardless of whether we agree on how many several is or have an idea of what kind of animals were in the road, our ability to make use of that information as actionable intelligence depends upon how much we trust that information.

16.2.2 Uncertainty About the Content

In the preceding section we used the following sentence as an example of ambiguity:

4) Sally gave Mary her book.

While our previous confusion about the ownership of the book continues, we can assume this event (the handing over of a book) actually took place. However, suppose the sentence read as follows:

6) It is possible that Sally gave Mary her book.

Now we are no longer certain as to whether indeed the event of Sally giving Mary a book occurred. There are multiple ways to view this. Perhaps there was no exchange of a book. Perhaps Mary did receive a book, but it was Georgina who gave it to her. Perhaps the players stayed the same, but it was Mary who gave Sally the book and not the other way around. The presence of "it is possible" changes the credibility of the event significantly. Natural languages are filled with a variety of different mechanisms which inject some uncertainty into the soft data they convey; analysis of these mechanisms will support fusion of soft data in that we may better assess the quality of the data which we are using.

16.2.2.1 "Words of Estimative Probability"

In his 1964 article, Sherman Kent of the United States Central Intelligence Agency relates the following anecdote about a conversation concerning an intelligence report on the possibility of a Soviet invasion of Yugoslavia:

- A few days after the estimate [NIE 29-51, "Probability of an Invasion of Yugoslavia in 1951"] appeared, I was in informal conversation with the Policy Planning Staff's chairman. We spoke of Yugoslavia and the estimate. Suddenly he said, "By the way, what did you people mean by the expression 'serious possibility'? What kind of odds did you have in mind?" I told him that my personal estimate was on the dark side, namely, that the odds were around 65 to 35 in favor of an attack. He was somewhat jolted by this; he and his colleagues had read "serious possibility" to mean odds very considerably lower. Understandably troubled by this want of communication, I began asking my own colleagues on the Board of National Estimates what odds they had had in mind when they agreed to that wording. It was another jolt to find that each Board member had had somewhat different odds in mind and the low man was thinking of about 20 to 80, the high of 80 to 20. The rest ranged in between [8].

What makes this anecdote particularly pertinent is that the individuals with whom Kent spoke were all intelligence analysts, that is, people who were working in the same domain (intelligence), who most likely had similar educational backgrounds and similar job training. It might be reasonable to assume people working in the same domain with a similar training and educational backgrounds would have some consistent understanding of the words and phrases routinely used within their working environment; however, this anecdote shows us that the understanding of such terms can be quite diverse, even in a relatively homogeneous group.

Intrigued by Kent's observations, CIA analyst Richards Heuer [9] ran an informal study in which a number of his CIA colleagues were requested to assign a single probability to some twenty-five common uncertainty expressions used by the analysts. While the probabilities assigned by the analysts to some of the terms were clustered very closely (better than even, about even, highly unlikely), there were several which varied quite dramatically: highly likely ranged more than 40 percentage points, as did improbable, probably not and chances are slight, while the range for probable was from 25% to just over 90%.

A couple of decades later, Rieber [10] requested analysts in training at the CIA's Kent School (named after Sherman Kent) to assign ranges of percentages rather than single values to a smaller number of hedges. Similar to Heuer's informal study, the ranges of percentages vary from quite narrow to relatively large, but the ranges are not necessarily identical to those in the first chart, even for identical hedges. One can almost assume that giving the task of assigning probabilities for hedges to any random group of English speakers will result in somewhat different numerical ranges.

In the half-century since Kent's initial work, the US intelligence community continued to struggle to standardize the terminology which they used to assess situations, in order to reach a common understanding of the meaning of those terms. Ultimately, the intelligence communication settled on a standard spectrum of words of estimative probability: remote, very unlikely, unlikely, even chance, probably/likely, very likely, almost certainly.

The discussion above involves examples of uncertainty which are straightforward and immediately obvious to the reader. However, there are other, less obvious forms of uncertainty which are less obvious, often overlooked or ignored; we will examine these in the following sections.

16.2.2.2 Hedges and Other Evidential Markers

In general, when asked to consider markers of uncertainty in natural language utterances, the first group of words that comes to mind are the expressions (mostly modal adverbs) such as we have seen in the preceding subsection: possibly, probably, likely, etc. The next categories are often modal verbs: might, could, may, etc., followed by nouns such as likelihood, possibility, probability, and so on. Lexical verbs such as suggest, assume, seem, guess, etc., likewise convey uncertainty, as do adjectives such as possible, probable, doubtful, etc. The elements above are manifestations of uncertainty are generally included in a group of lexical markers called hedges.

The term "hedge" is attributed to Lakoff [11] to mean any lexical or grammatical form which indicates "fuzziness" in natural language. Inspired by the mathematical theories of Zadeh, Lakoff defines a broad range of lexical and grammatical elements in natural languages which indicate any weakening of the formulation of propositions, which express vagueness or imprecision:

- For me, some of the most interesting questions are raised by the study of words whose meaning implicitly involves fuzziness – words whose job is to make things fuzzier or less fuzzy. I will refer to such words as 'hedges' [11].

Since Lakoff's first article, the definition of hedging which he proposed has shifted to focus more narrowly on expressions of uncertainty or commitment on the part of the speaker. Some researchers consider modals verbs (could, should, might, etc.) to be hedges, others classify them differently (see Rein, Ref. [4], for a deeper discussion).

But as we have seen at the beginning of this section, hedges are not the only elements which signal uncertainty in text information. There are markers that indicate that where the informational contained in the sentence comes from when it is not a direct observation. These are called "evidential markers" which can be divided into two broad categories: hearsay and what Bednarek [3] refers to as "mindsay".

Hearsay; that is, information which the writer has acquired from another source (not himself), is uncertain by nature in that we can never be certain that the writer has correctly and fully understood or recorded what the original source said (or indeed if there is only a single hearsay source rather than a chain) or that the unrelated context would cause us to view the information differently. As a result, we cannot be certain that the information that has been passed on is reliable.

Mindsay is information which comes not from a secondary or tertiary (external) source, but from the primary source and which is based upon belief, speculation, assumption rather than direct observation; that is, it is a product of some process in the primary source's mind.

Take, for example, the following sentences:

7) John is a terrorist.
8) The CIA has concluded that John is a terrorist.
9) I believe that John is a terrorist.
10) Mary thinks that John is a terrorist.

In each of these sentences, the relationship pattern ("content") of the sentence might produce the relation John IS-A terrorist. In sentence 7) there are no lexical clues to indicate where the information came from, nor how

credible the speaker considers the information to be. Thus, the basis of our decision as to whether John is, in fact, a terrorist must come from somewhere else, for example, previous knowledge of John's activities or from our belief in the speaker's truthfulness.

However, there are lexical clues contained in three of these four sentences which give us reasons to doubt on some level whether John is terrorist. In sentence 8) and sentence 10) there are clear indicators of third-party information, i.e., hearsay, which may or may not have been repeated accurately by the writer.

In sentence 9) and sentence 10) there is an indication of belief, i.e., mindsay, rather than knowledge on the part of the sources. In sentence 9) it is clear this is a first-person reporting of the speaker's belief. Sentence 10) is particularly interesting because it is, in fact, ambiguous. In one interpretation, one could say that it contains both hearsay and mindsay: the writer informs us about something another person (Mary) has told him about her thoughts regarding John. A second interpretation could be that the writer expresses his own belief (mindsay) about what Mary thinks (likewise mindsay). Regardless of the interpretation, the strength of the assertion of John being a terrorist is weakened.

A single sentence may contain multiple clues as the veracity of the main proposition of the statement. For example, consider the variations on sentence 7):

11) The CIA has concluded that John is probably a terrorist.

12) The CIA has concluded that John is most probably a terrorist.

In sentence 11), inserting the adverb *probably* into sentence 8) weakens the assertion of John being a terrorist, whereas in sentence 12) adding *most* before probably strengthens the assertion as opposed to sentence 11), but it still remains weaker than in 8). If requested, an English speaker would be able to identify and rank assertions from strongest to weakest according to the lexical clues the writer has left in the sentence.

Other factors that may be considered in the assessment of the strength or weakness of an uncertain proposition include variation such as whether in hearsay the original source is named (the CIA) as opposed to an unnamed source (*rumor has it*) or general knowledge (*it is widely accepted*). The relative strength of mindsay may also be determined via the verb used, e.g., *inferred* is stronger than *guessed.*

Since most, if not all, decision-making models using information will use some sort of mathematical weighting system based upon the perceived certainty or doubt about the veracity of the data which populates the model. Frajzyngier [12] comments, "the different manners of acquiring knowledge correspond to different degrees of certainty about the truth of the proposition." Models designed for device-based information, such as sensors, cameras, radar, etc., may use factors based upon testing, calibration, and the influence of environmental factors, such as light, heat or humidity, to adjust the reliability of readings or to fine-tune results. The human-generated information, in contrast, which comes in as text or speech is often assessed by a human, who uses her understanding of various factors including the background knowledge domain of the information, heuristic or scientific models, or even just a "gut feeling" to evaluate the information and assign it some sort of credibility weight. In lieu of other information, lexical markers may also play a role in the assessment; the analyst may well assign a lower "truth value" to information tagged by the informant through hedges to be *doubtful* than to information considered as *highly likely*.

Whereas many hedges such as the words of estimative probability discussed in the preceding subsection tend to be relatively easy for humans to assign a numerical weight (for an overview, see Ref. [5]), there is less research to be found on numerical weights for hearsay and mindsay markers. One could easily argue that weighting of the information from different types of sources in such a hierarchy is implicit. For example, direct perception (e.g., *I saw*) is often considered more reliable than hearsay (*he told me*); several authors including Goujon [13], Marin-Arrese [14], and Liddy *et al.* [15] have looked at such relative values.

16.2.2.3 Passive Voice, Depersonalization, Time, etc.

Hedges and evidential markers are quite obvious indicators of uncertainty, even to non-linguists. However, there are some more subtle ways in which uncertainty may appear.

In his discussion on hedges, Hyland [16] includes several other phenomena such as passive voice, conditionals (*if* clauses), question forms, impersonal phrasing and time reference. Particularly in scientific writing, the use of passive voice and impersonal phrasing are widely, almost universally, used, conveying an undertone of "but I might be wrong or have overlooked something." With regard to impersonal phrasing, Hyland writes:

> *...the writer inevitably uses a wide range of depersonalized forms which shift responsibility for the validity of what is asserted from the writer to those whose views are being reported. Verb forms such as argue, claim, contend, estimate, maintain and suggest occurring with third person subjects are typical examples of forms functioning in the way, as are adverbials like allegedly, reportedly, supposedly and presumably.*

Passive voice and impersonal phrasing, however, can also be used to express politeness, rather than uncertainty, which can only be determined by knowing some information about the context of the statement. Likewise, passive voice and impersonal phrasing can also sometimes used in instances of differences in social ranking or power, in order not to offend. In such cases, these forms are not intended to create doubt about the veracity of the proposition, but to soften the impact of a message on the intended audience.

While time might not immediately spring to mind when considering expressions of uncertainty, it nevertheless plays a significant role, and should therefore be discussed.

Any sentence which is formulated in the future tense is inherently uncertain, simply because the event or state which is described has not happened yet:

13) Mary will be at the meeting next week.

That is, of course, unless Mary decides not to go for some reason, her plane flight is cancelled due to snowfall, or she gets sick and lands in the hospital, or worse.

That being said, some future things are more certain than others:

14) The next presidential election in the US will take place in November 2020.

Clearly, unless something unbelievably catastrophic happens, there is virtually no chance that the elections in the United States will not take place in the month and year named because of legal requirements in the voting laws, so this may be treated as a fact, rather than a possibility.

In other cases, it is a bit less clear. Take the case of routine, recurring behavior:

15) Sam always attends the meetings of local political action group every Tuesday at 7 p.m.

One can expect to find Sam at this meeting every Tuesday – unless, of course, Tuesday is a holiday and the meeting is cancelled, or unless Sam, like Mary in sentence 13) has something come up to prevent his attendance. In other words, recurring behavior may be a good indicator of future behavior, but not a guarantee.

For intelligence purposes, information based upon future actions often plays a very significant role, particularly in preventative measures, but should nearly always be considered uncertain, until the expected date of that action has passed; at that point the event has either occurred or not occurred, and an update should be made to the knowledge base which is being used.

Now that we have examined a number of ways in which formulations may represent uncertainty about the veracity of the informational content in natural language, in the following section we will examine how we might exploit these to algorithmically generate initial credibility weights.

16.3 QUANTIFYING EVIDENTIALS FOR CREDIBILITY WEIGHTING

> *When one admits that nothing is certain one must, I think, also admit that some things are much more nearly certain than others.*
>
> Bertrand Russell [17]

As previously discussed, when humans communicate with one another they transmit content information, but this content is often surrounded by additional, non-content information from the speaker intended to convey the speaker's stance to that information.

Text analytics to extract actionable information from text utilize algorithms to locate identify certain patterns which may identify objects or individuals, events, relationships and other useful information. The weakness of these algorithms is that they seldom, if ever, take into account that some of those patterns may be couched in language that indicates those identified events or relationships may be questionable or false. In other words, for all intents and purposes the information extracted using text analytics is treated as "fact," even though there may be clear evidence that they are not.

In this section, we take a look at some of the work which has been done thus far in assigning numerical values to evidential expressions. As we will see below, this earlier work tends to focus almost exclusively on certain types of hedges, in particular expressions containing modal verbs (*should*, *could*, *might*), other verbs indicating conviction or another source (*believe*, *doubt, according to*, *assume*, *guess*), adverbs (*possibly*, *probably*, *likely*), adjectives (*it is possible*, *probable*) and some nouns (*small chance*, *high likelihood*). In most of these studies, attempts have been made to ascertain numerical "values" for the various expressions, in general by asking participants in a study to locate the expression along a scale, from, say, 0 to 100. Furthermore, hedges are often strengthened or weakened by the use of boosters and downtoners (*very likely*, *rather improbable*), requiring adjustment to their assigned values.

Other types of hedges, for example, those dealing with informational source such hearsay, conjecture, inference, etc., may often be ranked relative to each other in a hierarchical sense, but there appear to have been no attempts to assign numerical values to the evidentials in this category.

To make things even more complex, hedges do not always appear alone in a sentence: *I believe it is possible that Mary could be*…; Clausen [18] found "that uncertain sentences often contained multiple hedge cues, sometimes up to 4 or more." Therefore, we often must try to assign a weight to the proposition which is based upon the interaction of multiple hedges.

One solution is to determine, in advance, all possible combinations of hedges, boosters and downtoners and assign them individual values. This would be a brittle solution, broken as soon as a combination does not appear in the table of values. Luckily, Crompton [19] points that "compounding of hedges is quite common, but the elements of each compound are still distinguishable"; the reader can easily corroborate Crompton's assertion in the example from the preceding paragraph (*I believe it is possible that Mary could be*…). Recognizing this provides us with a basis to support a more robust solution, which is to determine the weights assigned to individual hedges and assign a composite weighting.

However, as we will see, ultimately the numerical values per se are irrelevant, in so far as there are no definitive "universally true" numerical values assigned to any of these hedges. Indeed, in the examples that follow various researchers have used their own different (arbitrary) weighting scales, resulting different weight values and ranges for the same hedges.

What does, however, appear to be "universally true," as we shall see in the following sections, is the general ordering in which humans tend to organize the various hedges with their accompanying boosters and downtoners. This is of particular significance to us, as it allows us to assign "relative" weighting while at the same time freeing us from being tied to a specific mathematical weighting system. In other words, it allows us the freedom to assign, for example, single evidential value (crisp weights) or a range of values (fuzzy weights) to any given hedge or chain of hedges depending on the underlying application which is being used.

16.3.1 A Brief Look at Studies Assigning Values to Words of Estimative Probability

In Section 16.3.2.1 we discussed the informal exercise in which CIA analyst Heuer [9] requested a number of colleagues to assign a single probability to a number of commonly used hedges. Figure 16-2 shows the hedges along with a mapping of the various probabilities assigned to each hedge.

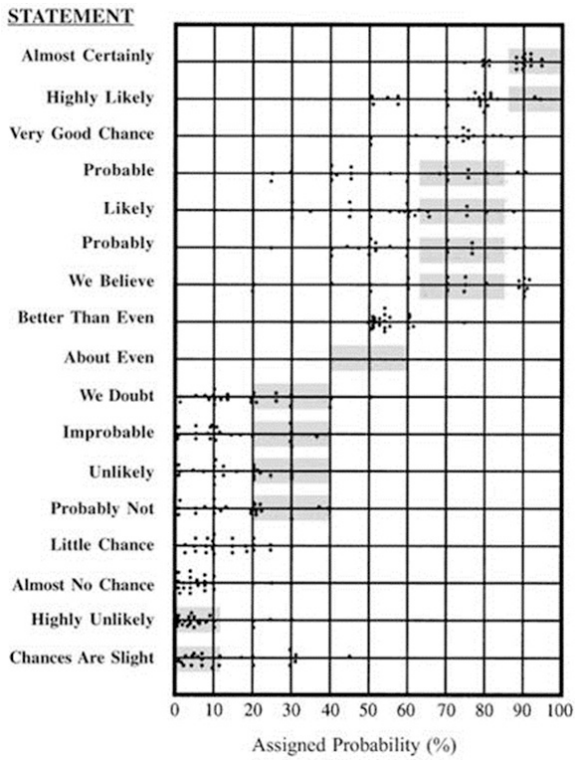


Figure 16-2: Probabilities Assigned by CIA Analysts to Various Hedges [9].

The probabilities assigned to a number of the hedges were clustered very closely (*better than even*, *about even*, *highly unlikely*). A number varied quite dramatically: *highly likely* ranged more than 40 percentage points, as did *improbable*, *probably not* and *chances are slight*, while the range for *probable* was from 25% to just over 90%.

Staying within the analyst realm, Rieber [10] requested analysts training at the Kent School (named after Sherman Kent) to assign ranges of percentages instead of specific values to a number of hedges. The results are shown in Figure 16-3.

Again, one can see that the ranges of percentages range from quite narrow to relatively large, but the ranges are not necessarily identical to those in the first chart, even for identical hedges (compare *probable* in both). One can almost assume that giving the task of assigning probabilities for hedges to any random group of English speakers will result in somewhat different numerical ranges.

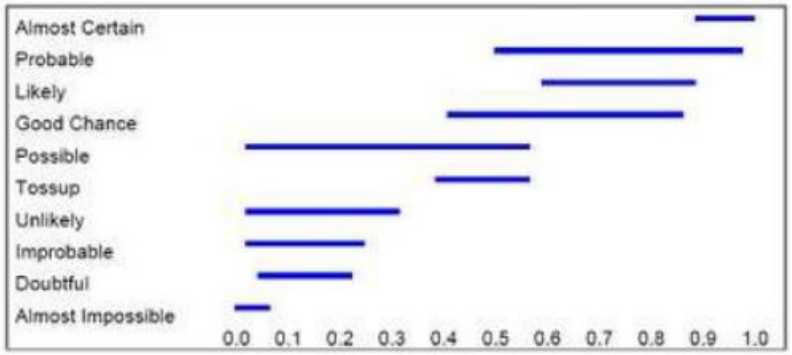


Figure 16-3: Ranges of Percentages Assigned to Hedges by Analysts in Training [10].

Brun and Teigen [20] investigated the numerical weights of probability expressions in several different contexts: a discussion of medical treatment effectiveness (conversations between paediatrician and parents of sick children), opinions on current events, and usage in videotaped television news reports. Their focus was on the evaluation of not only the weighting of differences between various domains (medicine, news, opinion columns), but in the case of the medical discussion, the differences between the understanding of the expressions between the doctors and the parents of the sick children). The values assigned are shown in Figure 16-4.

PROBABILITY RATINGS (0 TO 6 SCALE) AND PERCEIVED AMBIGUITY OF PROBABILITY WORDS IN A MEDICAL TREATMENT CONTEXT (STUDY II)

			Treatment context					
			Physicians			Parents		
	No context (students)							
Expression	Mean	*SD*	Mean	*SD*	Perceived ambiguity (%)	Mean	*SD*	Perceived ambiguity (%)
Impossible	0.1	0.4	0.0	0.2	96	0.1	0.2	93
Improbable	1.2	1.1	0.6	0.5	87	0.3	0.5	83
Doubtful	1.3	0.7	1.0	0.4	92	0.9	0.5	88
Perhaps	2.9	0.8	2.3	0.9	73	2.4	0.9	73
Possibly	3.1	1.0	2.2	0.9	73	2.5	1.1	70
Chance for	3.1	0.9	2.3	1.0	77	2.7	0.9	67
Danger for	3.3	1.3	1.5	0.9	76	1.1	1.0	75
Possible	3.6	0.8	2.5	0.9	79	2.9	0.9	74
Assumedly	4.1	1.0	3.4	1.1	65	3.5	1.2	69
Good hope	4.2	0.9	3.9	0.9	84	4.4	1.0	77
Likely	4.2	0.9	3.9	0.9	76	3.5	1.0	69
Good chance	4.5	0.6	4.2	0.8	86	4.5	0.9	79
Probable	4.6	0.8	4.4	0.8	73	3.8	1.1	66
Small doubt	4.6	1.2	4.8	0.9	64	4.1	1.9	54
Mean	3.2	0.9	2.6	0.8	79	2.6	0.9	74

Figure 16-4: Weights Assigned to Expressions of Uncertainty Used in the Context of Medical Discussions Between Paediatricians and the Parents of Sick Children. Brun-Teigen [20].

Renooij and Witteman [21], whose interest in the quantification of probabilistic expressions comes from the field of computer (Bayesian) modelling in medicine, evaluated three groups: medical students, other students, and the first two groups combined. From the information gathered, they created the simplified probability scale shown in Table 16-1.

Table 16-1: Final Scale with Seven Categories of Probability Expressions Plus Their Calculated Probability Points [21].

	Expression	Probability (%)
I	Certain	100
II	Probable	85
III	Expected	75
IV	Fifty–fifty	50
V	Uncertain	25
VI	Improbable	15
VII	Impossible	0

The above are just a handful from the many studies done by linguists and other researchers in papers such as Wesson and Pulford [22], Ayyub and Klir [23] and others in which single words or multi-word hedges have been evaluated and given numerical weights (probabilities, odds, etc.) by test subjects.

While there are variations among the studies in the values assigned to any of the expressions, from which one can draw the following conclusion: there are *no universally accepted* values. For a deeper discussion, see Ref. [4].

While the expressions discussed in this subsection are relatively easy for people to assign some sort of numerical value, there are other indicators of uncertainty which are not so easy to quantify. We examine these in the following subsection.

16.3.2 Relative Weightings of Other Evidential Markers

As we have seen above, speakers of English may assign different numerical values but still consider *probably* and *unlikely* to be more or less at opposite ends of a probability range. Thus, regardless of numerical values assigned, when taken as a group, what can be seen is that these elements may be relatively ordered along a scale from stronger to weaker (or higher to lower; or more true to less true; to name just a few possibilities).

For example, in general, English speakers would agree to the following ordering:

$$unlikely < probable$$

even if they do not agree on the precise numerical values which they assign to these two words of estimative probability.

Based upon the overall consistency of such relative weightings, the US intelligence community settled on a standard spectrum of such expressions, as shown in Figure 16-5.



Figure 16-5: Words of Estimative Probability.

Note: Source: Graphic displayed in the 2007 National Intelligence Estimate "Iran: Nuclear Intentions and Capabilities," as well as in the front matter of several other recent intelligence products (via Friedman/Zeckhauser [24]).

As seen in Figure 16-5, very often these expressions are modified by other words, which can strengthen or soften their original meanings:

highly unlikely < unlikely < probable < very probable

Negation, of course, has a dramatic effect on the ranking from weaker to stronger; we will discuss this in more depth later in this chapter.

But, as discussed previously, it is not simply the use of words of estimative probability which are indicative of uncertainty about the validity of information, but also lexical markers for hearsay and mindsay.

For example, in general, English speakers would agree to the following ordering:

I saw > I infer > my neighbor told me

As mentioned previously, a number of researchers (Goujon [13] Marin-Arrese [14] etc.) have examined the relative strength or weakness of a proposition based upon such markers. DeHaan [25] proposed a cross-linguistic comparison of source evidentiality which reflects the previous ordering:

sensory > inferential > quotative

While there has been research by linguists on this topic, there has been little attempt to assign any sort of (numerical) uncertainty weighting to these evidential markers, possibly in part because it seems to be a non-obvious exercise. One can, however, argue that there is an implicit weighting based upon this hierarchy. There appears to be consistency in the rankings between different groups of people surveyed on these topics as documented by research, from which we can conclude that there seems to be some sort of universal scalar for the various elements which we may exploit for our purposes.

It should be clear from the discussion above that the assignment of numerical values (probabilities, odds) to the lexical and grammatical elements which are of interest to us is not easy. However, it can be very useful to assign uncertainty weights to propositions based upon these clues, especially when fusing uncertain information to use as the basis for informed decision making.

There remains one more complication, namely the fact that humans do not always use these elements in isolation. Rather, it is not uncommon that several different markers appear in a single sentence:

16) I believe John told me that it is very possible that Mary will arrive on Sunday.

How certain can we be that Mary will indeed arrive on Sunday, based upon this sentence? While different readers may, if requested, assign different probabilities to her arrival, in general one can say that each lexical marker (mindsay, hearsay, words of estimative probability) in this sentence collectively increases our uncertainty.

Natural languages are very flexible, allowing for an infinite combination of words. Thus, listing all possible combinations of these lexical markers and assigning each combination a value would be an arduous (and probably pointless) task, to say the least.

However, in Ref. [5] we have shown that it is possible to exploit certain types of lexical items to evaluate this chaining. For example, intensifiers may be used to weaken (downtoners) or to strengthen (boosters) the evidential weight of elements such as adverbs or adjectives. That is, use of the downtoner *somewhat* weakens *likely* in *somewhat likely*, and similarly the booster *very* will turn *likely* into the stronger *very likely*. If asked to arrange the resulting terms in order from weakest to strongest, speakers of English will generally arrive at the following relation:

somewhat likely < likely < very likely

Not unexpectedly, there is the reverse effect when we use *somewhat* and *very* with the modal adverb *unlikely*:

very unlikely < *unlikely* < *somewhat likely*

Figure 16-6 shows a relative placing of a proposition ***p*** modified by the above-mentioned hedges along a scale from ***p*** *is untrue* to ***p*** *is true* for any given proposition ***p***.

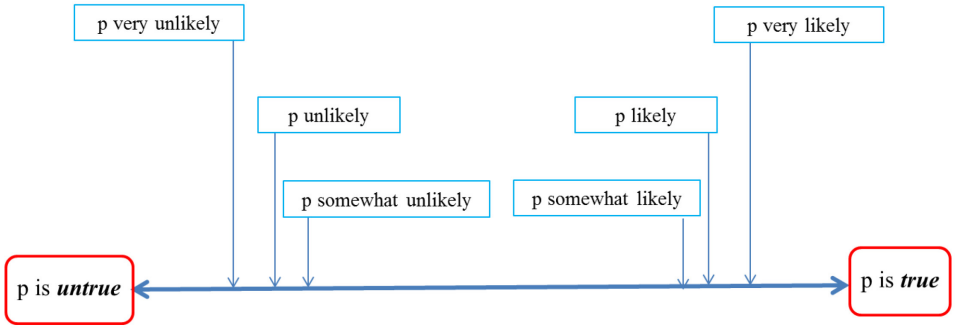


Figure 16-6: Ranking of *Unlikely* and *Likely* Modified by Booster *Very* and Downtoner *Somewhat* for a Proposition *p*.

But in both cases, we can say that the addition of *very* increased our certainty about a proposition ***p*** being either true or false: if something is *very likely*, we are pretty certain it is true (or will happen); if something is *very unlikely*, we are quite certain that it is **not** true (or will **not** happen). When we are truly uncertain – the coin is still in the air – we cannot say we are more or less certain to believe ***p*** to be true or untrue, and are therefore stuck in the middle between ***p*** being true and ***p*** being false (Figure 16-7).

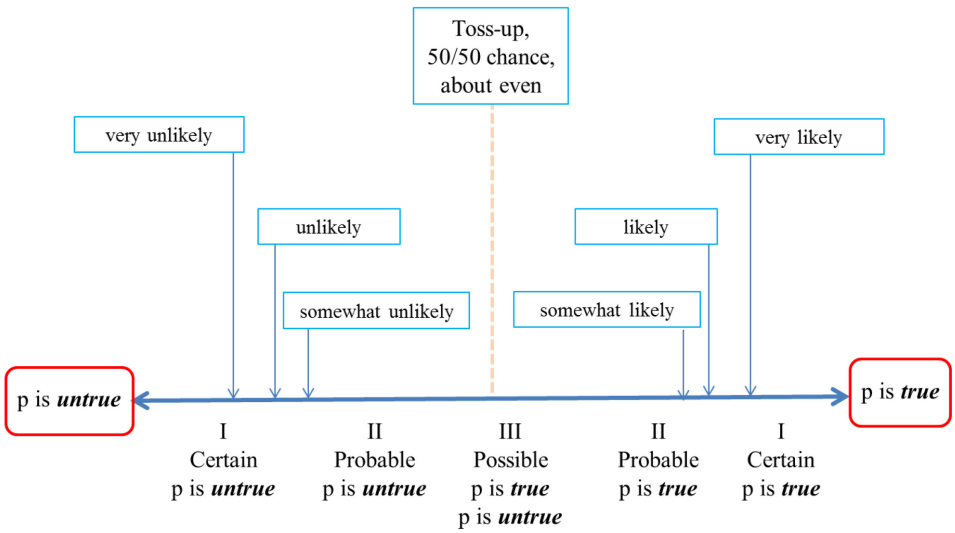


Figure 16-7: Some Modifications to Figure 16-6, Including Labels and Expressions for Complete Uncertainty.

While many people view uncertainty on a scale ranging from uncertain to certain (i.e., equivalent to a "0 – 100 scale", with 0 representing "uncertain" and 100 "certain"), it turns out that the scale is bipolar: the point of maximum uncertainty lies in the middle of the scale, while maximum certainty lies at both ends of the scale, as illustrated in Figure 16-8:

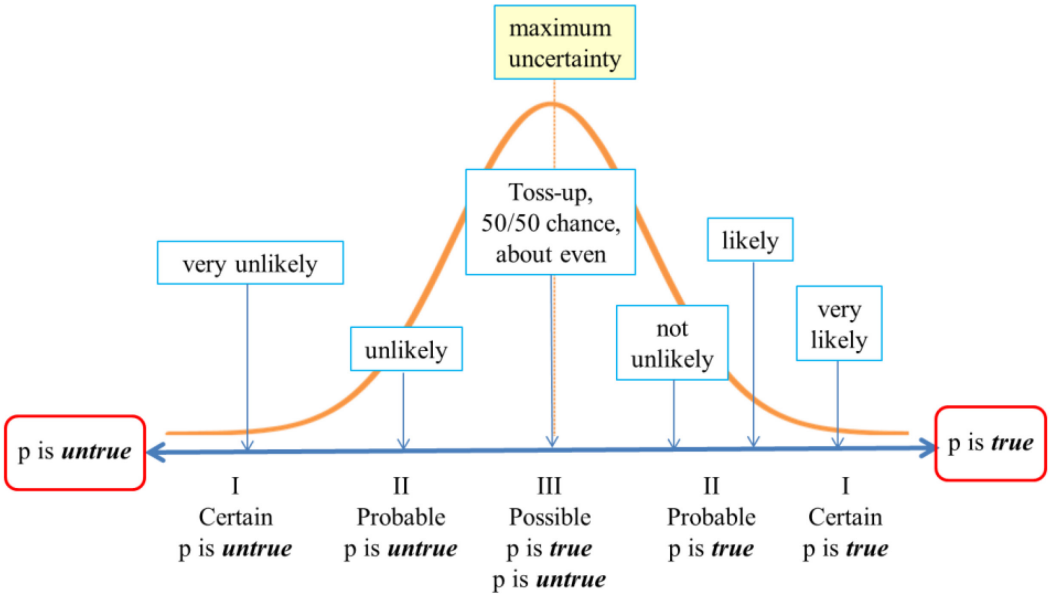


Figure 16-8: Bipolar Scale Based on Showing Point of Maximum Uncertainty in the Center.

Using this bipolarity as a basis, we can define a numerical scale in which the midpoint (the point of maximum uncertainty) is zero, while the end of the scale representing absolute certainty that ***p*** is true is assigned the value 1.0 and the end of the scale representing absolute certainty that ***p*** is untrue is assigned the value -1.0, as shown in Figure 16-9.

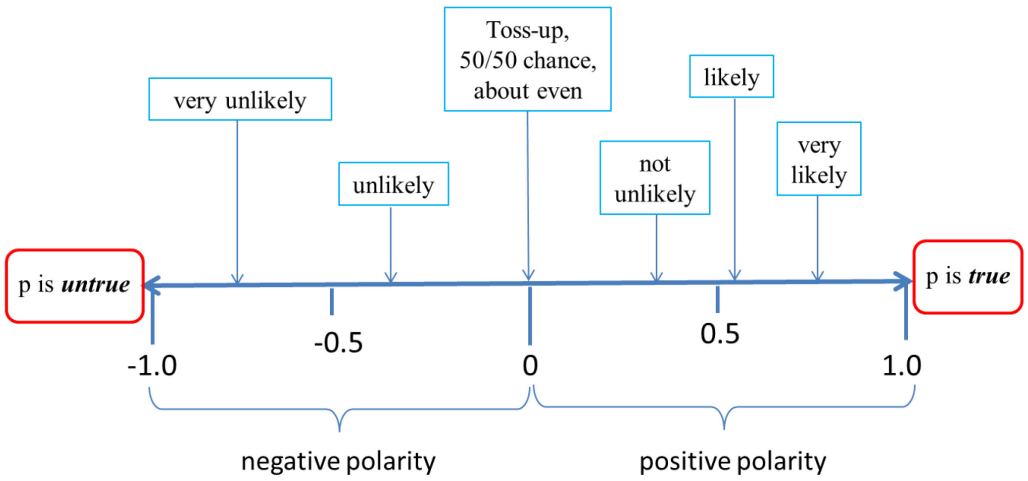


Figure 16-9: A Numerical Scale for Certainty *p* is Untrue (-1.0) to Certainty *p* is True (1.0), with the Point of Maximum Uncertainty Assigned the Value Zero.

We can then exploit this scale to help us to automatically determine relative evidentiality weights for chained and modified evidential markers. For this, we define the effect of a modifier on the original weight as follows:

$$w_{modifiedhedge} = w_{original} + p_{original} * effect_{modifier} * (1 - | w_{original} |) \quad (16\text{-}1)$$

where $p_{original}$ is the polarity of the original hedge.

For example, suppose we assign the weight $w_{likely} = 0.6$ to *likely*, and the weight $w_{unlikely} = -0.6$ to *unlikely*. If we have determined for our model that the adverb *very* amplifies (strengthens) that which it modifies by 0.3, then we can use the formula to obtain:

$$w_{very_likely} = 0.6 * + (1) *(0.3) * (1 - | 0\text{-}6 |) = 0.72 \quad (16\text{-}2)$$

$$w_{very_unlikely} = 0.6 * + (-1) *(0.3) * (1 - | -0\text{-}6 |) = -0.72 \quad (16\text{-}3)$$

which places, as expected, *very likely* to the right of *likely* on the scale, and *very unlikely* to the left of *unlikely*, as they appear in Figure 16-6 above.

Negation can be affected by simply modifying the polarity, as shown in Figure 16-10. Note that negating a negatively poled expression results in something less certain than its positively poled antonym, i.e., whereas *not likely* and *unlikely* are usually considered to be more or less equivalent, *not unlikely* is generally considered to be weaker than *likely* [26]. (The reader should note that the issue of negation is a significant research topic in linguistics as natural languages do not always reflect the black-and-white situation of pure logic, particularly in the area of negation of negatives, but rather offer subtle shadings.)

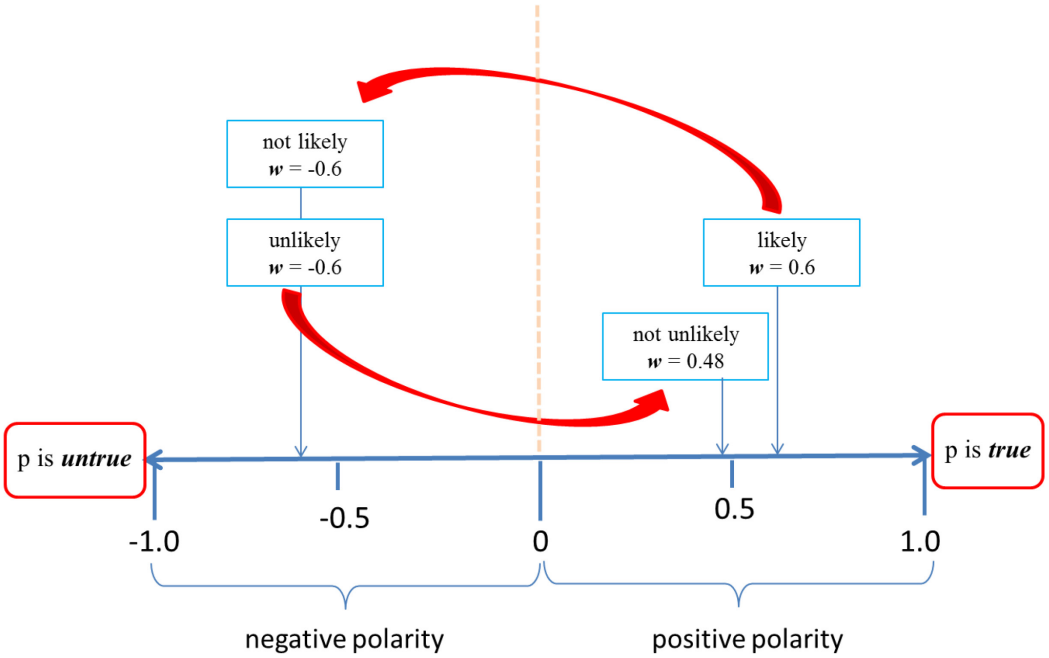


Figure 16-10: Negation of Words of Estimative Probability.

Similar to words of estimative probability, as discussed above, lexical clues indicating hearsay or mindsay may be assigned values and modified, as described above. A sample is shown in Figure 16-11.

Thus, by evaluating the various multiple lexical expressions surrounding the content, we can come up with a rough estimation of the credibility of the information based upon clues the reporter has left us in the communication.

Once we have the ranking, we can map the results to existing scales such as a fuzzy scale using words of estimative probability or to a numerical scale such as percentages. Two simple examples appear in Figure 16-12.

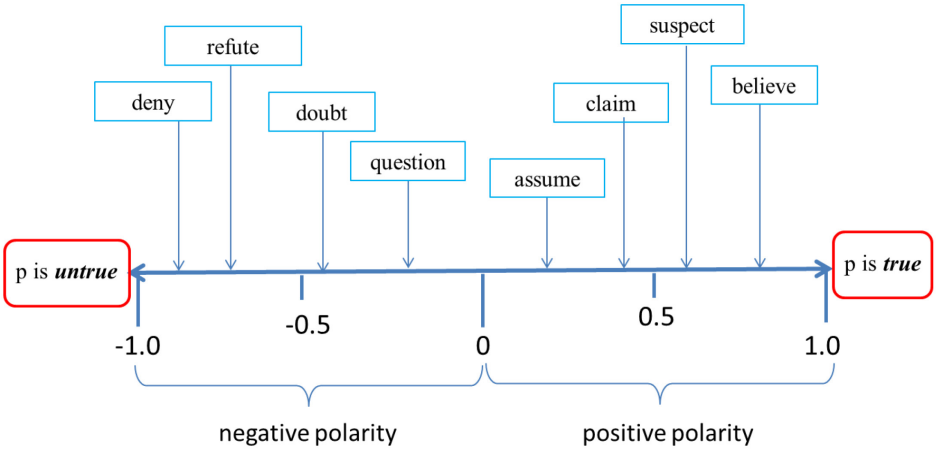


Figure 16-11: Example of Relative Weightings of Various Verbs Expressing Uncertainty about Propositional Information.

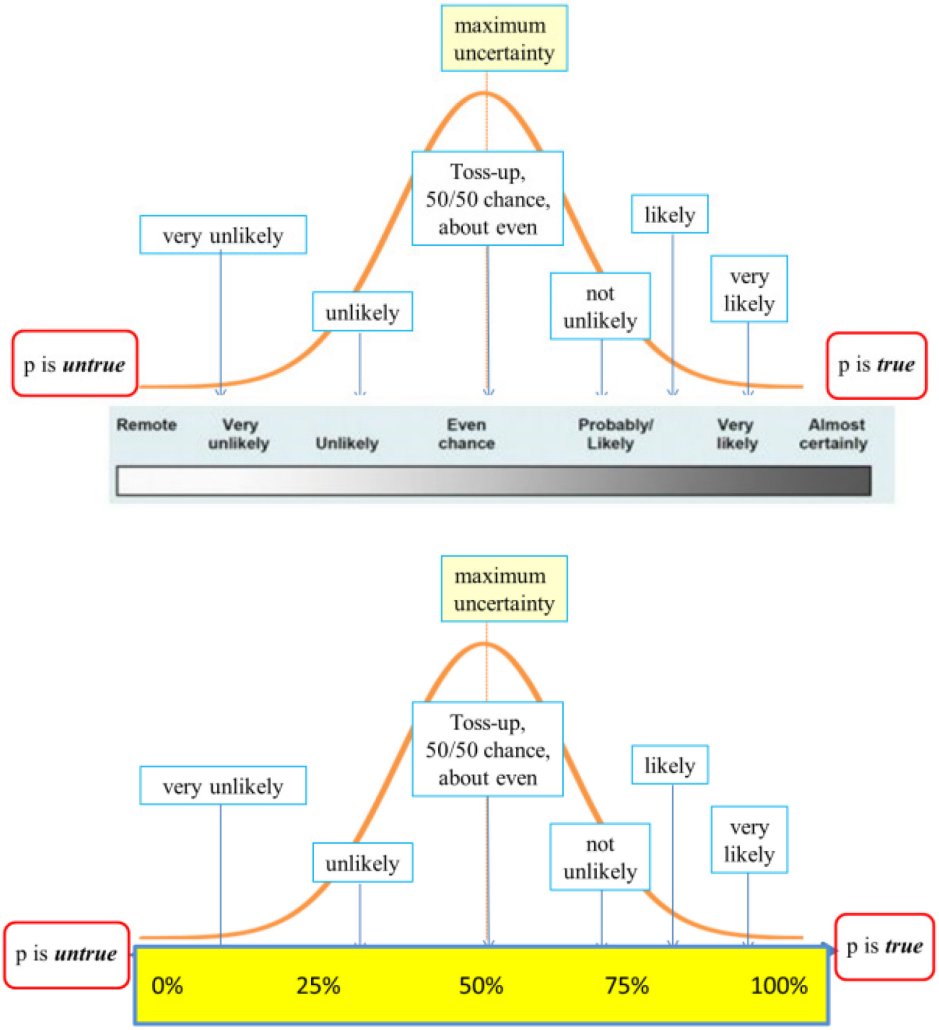


Figure 16-12: Examples of Mappings of Relative Rankings onto a Scale of Words of Estimative Probability (Top) and Percentages (Bottom).

It must again be reiterated here that the values assigned to expressions, as well as the values to modifiers are defined by the user; as mentioned earlier, there are no universal values, but there is a certain consistency in relative understanding of the values (meanings) of various expressions.

For a more detailed, in-depth discussion of the underlying research as well as a more detailed discussion of the algorithms, please refer to Ref. [4].

16.3.3 A Few Examples

Before we take a look at a few simple examples, it must again be reiterated here that the values assigned to expressions, as well as the values to modifiers are defined by the *implementer*; as mentioned earlier, there are no universal values, but there is a certain consistency in relative understanding of the values (meanings) of various expressions. Exactly how the implementer views the relative strength or weakness of a given element may be dependent on a given domain (i.e., subject matter experts, such as doctors, use terminology with a similar intent within the field, but the use may not be understood in the same way by others, e.g., patients). Thus, for our examples, we will be assigning values which we find, for our purposes, to be useful.

So, with that caveat in mind, let us return to some of the examples above.

6) It is possible that Sally gave Mary her book.

The informational content of this statement is Sally giving Mary the book. The writer gives us "it is possible" to indicate there is some doubt. As we have seen above, participants asked to assign a numerical value to *possible* are able to do so. We have decided that *possible* is a relatively weak hedge, that it lands in the positive range (0 – 1.0) of the bipolar model above and we assign it the value 0.3. Using the mapping examples from Figure 16-12 that would translate into the "probably/likely" of the Words of Estimative Probability (WEP) scale, and around 65% in the percentage scale. For most readers this is probably something they could agree upon as being reasonable (even if each reader might personally assign a slightly different value along the range).

If we add the booster *very* to *possible* in sentence 6) as in sentence 17), we end up with a stronger statement, one which moves us closer to 1.0 on the bipolar scale. As described in the preceding section, we can assign a numerical value to the incremental effect of *very* on the informational content.

17) It is very possible that Sally gave Mary her book.

Again, this will generally be intuitive for the reader, and indeed, we see this reflected in the various works of researchers looking at the effects of hedging, as discussed at length in the section above.

However, if we look at sentence 18) below, we have the same informational content as sentence 6) but a different indicator of uncertainty. Rather than a hedge, we now have an example of mindsay. While most readers will accept the notion that *believe* is less certain than *know*, assigning a numerical value to *believe* may not be an easy task. However, in order to indicate that the informational content is uncertain, for our purposes, we have decided that the use of *believe* as a marker of mindsay weakens the certainty of informational content by 30%. Thus, if we start with informational content of 1.0 (perfect knowledge) on our bipolar we could say that believe put us at 0.7 on the bipolar scale and the mappings to the WEP and percentage scales in Figure 16-12 would be similar to those discussed for 6) above (but slightly stronger).

18) I believe that Sally gave Mary her book.

The reader may well ask why the use of simple look-up tables would not be an easy and convenient way of assigning values. If we return to statement 16) which appeared earlier in this chapter, we can observe a very common phenomenon, namely, chaining of various manifestations of uncertainty:

16) I believe John told me that it is very possible that Mary will arrive on Sunday.

In this statement we have mindsay ("I believe"), hearsay ("John told me"), a hedge ("possible"), a booster ("very") and future events (which are inherently uncertain until/unless they actually occur). Utilizing the values and effects which we have discussed in the preceding example, including weakening effects to hearsay of and to future events, we can calculate an evidential weighting for 16) as follows:

$$W_{evidential} = w_{possible} + p_{possible} * effect_{very} * effect_{mindsay} * effect_{hearsay} * effect_{future} * (1 - | w_{possible} |) \quad (16\text{-}4)$$

The ability to deal effectively with the chaining of multiple manifestations of uncertainty means that we are able to use a (relatively objective) methodology to calculate a reasonable evidential weighting for informational content based upon linguistic clues. This is of particular interest in text analytics, as during processing, a second pass can be made through the text available looking for indications of uncertainty which are an important factor in the determination of information quality.

16.3.4 A Note on the Numbers

It has been reiterated numerous times above that the values to be used in any such calculation are to be determined by the implementer, that there are no universally accepted values for any of the hedges, boosters or downtoners, let alone for hearsay or mindsay. For the latter two, as discussed Section 16.3.2, linguists such as DeHaan have found that there is tendency to order of lexical clues for non-hedging expressions: sensory, inferential and quotative; assigning a quantity to such terms may be difficult for most speakers of a language, and may be dependent on that individual's background knowledge about the informational content, belief system and other factors.

When I use the hedge *probably*, I am both letting you know that I am uncertain, as well as assuming that you will have a similar understanding of what *probably* means. However, the assumption is not guaranteed: as we have seen in Section 16.3.1 above, beginning with the anecdote from Sherman Kent through to Rieber's study some decades later (and many more between and after), there are broad differences in the understanding of and numerical weighting of various hedges. For any of those studies, if there had been one or two respondents more, the numbers associated with any expression of uncertainty, sometimes impressively calculated out to four places past the decimal, those numbers would be different. They are not immutable and should be understood as variable within the framework of any individual communication. However, as pointed out in Section 16.3, there tends to be a certain consistency in the relative rankings of terms.

However, this author often requested to provide some sort of "evidence" for the numbers used in the examples in the preceding subsection above, to which I can only reply, I chose based on my personal assessment of the meanings of the hedges, and a determination of relative effects of boosters and downtoners (i.e., I made sure that the effect of *very* is not as strong as the effect of *extremely* because in my native language of English, these two terms are understood in this way). I just need to stay within the general understanding of how boosters and downtoners function in the language I am working with.

Similarly, as an implementer, I may decide to introduce more granularity into the system, for example, to define two different hearsay effects, one for named secondary sources ("John told me") and another for unnamed ("rumor has it"). For the former, I may take into consideration an individual assessment of each hearsay source. If I determine John to be an unreliable source of information, I will discount his information more strongly than that of Mary, whose information is nearly always good. There may be disagreement between sender and receiver about the quality of a source: one party may consider *The Washington Post* to be a reliable source of, while another may consider the paper to be a spurious resource.

The purpose of the above was to propose a systematic methodology for looking at linguistic clues in natural language information to help us make an assessment of the informational quality of the content of a statement. In the original dissertation presenting this concept, the following quote appears at the beginning of the first chapter:

> *"You know what people are like," my father said. "Someone says, 'I suppose Leonard Kitchens could have put the rifle in the gutter, he's always in and out of the hotel,' and the next person drops the 'I suppose' and repeats the rest as a fact."* [27]

We further refer the reader also back to the quotation from Remy de Gourmont which appears at the beginning of this chapter: "… *in our daily life, we have less need of certainty than of a certain approximation to certainty."*

16.4 SUMMARY

In this chapter we have discussed the need for evaluation of uncertainty in natural language information to give decision makers a clear picture of the quality of that information. We discussed the various forms of uncertainty in natural language, categorized in two ways: uncertainty *within* the information which includes imprecision, vagueness, ambiguity, polysemy and so on, and uncertainty *about* the information, the constructs in which appear in text that tell us whether the information is reliable or not, and in which way. We then briefly present the outlines of a concept which allows us to automatically generate evidential weights for information derived through text analytics by examining lexical clues concerning original source or stance toward the veracity of the information which speakers embed in their communications.

16.5 REFERENCES

[1] De Gourmont, R. (1920). *Philosophic Nights in Paris*, 27. Boston, MA: J.W. Luce & Co.

[2] Gans, J.A., Jr. 'This is 50-50': Behind Obama's decision to kill Bin Laden. *The Atlantic*, Retrieved from https://www.theatlantic.com/international/archive/2012/10/this-is-50-50-behind-obamas-decision-to-kill-bin-laden/263449/ (October 10, 2012).

[3] Bednarek, M. (2006). *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*. London, UK: Continuum.

[4] Rein, K. (2016). I believe it's possible it might be so…: Exploiting lexical clues for the automatic generation of evidentiality weights for information extracted from English text. Universitäts- und Landesbibliothek Bonn, Germany. Retrieved from http://hss.ulb.uni-bonn.de/2016/4471/4471.htm.

[5] Gross, G.A., Nagi, R., Sambhoos, K., Schlegel, D.R., Shapiro, S.C., and Tauer, G. (2012). Towards hard+soft data fusion: Processing architecture and implementation for the joint fusion and analysis of hard and soft intelligence data. In: *Proceedings of Fusion,* 2012: 955-962.

[6] Anderson, S.R. How Many Languages are there in the World?, *Linguistic Society of America*. Retrieved from http://www.linguisticsociety.org/content/how-many-languages-are-there-world. (May 12, 2016).

[7] Claeser, D., Felske, D., and Kent, S. (2018). 'Token-level code-switching detection' using Wikipedia as a lexical resource. In: *Language Technologies for the Challenges of the Digital Age*, Rehm, G., and Declerck, T. (Eds.), GSCL 2017, Lecture Notes in Computer Science, Vol. 10713. Cham, Switzerland: Springer.

[8] Kent, S. (1964). Words of estimative probability. *Studies in Intelligence*, Fall 1964: 49-65.

[9] Heuer, R.J., Jr. (1999). *The Psychology of Intelligence Analysis*. Washington DC: Center for the Study of Intelligence.

[10] Rieber, S. (2006). Communicating uncertainty in intelligence analysis. https://www.qmdns.org/QMDNS2007/WebPages/PRES/IS3-1-Rieber.pdf.

[11] Lakoff, G. (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic,* 2:458-508. Dordrecht, Holland: D. Reidel Publishing Co.

[12] Frajzyngier, Z. (1985). Truth and the indicative sentence. *Studies in Language,* 9(2):243-254.

[13] Goujon, B. (2009). Uncertainty detection for information extraction. In: *Proceedings of the International Conference RANLP 2009*, 118-122. Borovets, Bulgaria.

[14] Marin-Arrese, J.I. (2011). Epistemic legitimizing strategies, commitment and accountability in discourse. *Discourse Studies,* 13: 789-797. Sage Publications. https://doi.org/10.1177/1461445611421360c.

[15] Liddy, E.D., Kando, N., and Rubin, V.L. (2004). Certainty categorization model. In *The AAAI Symposium on Exploring Attitude and Affect in Text. AAAI-EAAT 2004*, Stanford, CA: American Association for Artificial Intelligence.

[16] Hyland, K. (1998). *Hedging in Scientific Research Articles*. Amsterdam and Philadelphia: John Benjamins.

[17] Russell, B. Am I an atheist or an agnostic? (1949). *The Literary Guide and Rationalist Review*, 64(7):115-116.

[18] Clausen, D. (2001). HedgeHunter: A system for hedge detection and uncertainty classification. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, 120-125. Uppsala, Sweden: Association for Computational Linguistics.

[19] Crompton, P. (1997). Hedging in academic writing: Some theoretical problems. *English for Specific Purposes,* 16(4): 271-287. Amsterdam, Netherlands: Elsevier.

[20] Brun, W., and Teigen, K.H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes,* 41(3):390-404.

[21] Renooij, S., and Witteman, C.L.M. (1999). Talking probabilities: Communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning,* 22(3):169-195. Amsterdam, Netherlands: Elsevier.

[22] Wesson, C.J., and Pulford, B.D. (2009). Verbal expressions of confidence and doubt. *Psychological Reports,* 105(1):151-160.

[23] Ayyub, B.M., and Klir, G.J. (2006). *Uncertainty Modeling and Analysis in Engineering and the Sciences.* Boca Raton, FL: Chapman and Hall/CRC.

[24] Friedman, J.A., and Zeckhauser, R. (2015). Handling and mishandling estimative probability: Likelihood, confidence, and the search for Bin Laden. *Intelligence and National Security,* 30(1):77-99.

[25] De Haan, F. (2001). The place of inference within the evidential system. *International Journal of American Linguistics,* 67(2):193-219.

[26] Seright, O.D. (1966). Double negatives in standard modern English. *American Speech,* 4(2):124.

[27] Francis, D. (1997). *10 Lb. Penalty*, 206. London, UK: Pan Books.

Chapter 17 – UK AND US POLICIES FOR COMMUNICATING PROBABILITY IN INTELLIGENCE ANALYSIS: A REVIEW

Mandeep K. Dhami
Middlesex University
UNITED KINGDOM

David R. Mandel
Defence Research and Development Canada
CANADA

17.1 INTRODUCTION

Intelligence analysts are required to assess both current and future states of the world. For instance, what is North Korea's ballistic missile capability at the present time? How is its capability expected to develop over the next 5 years? And what is the strategic intent of these plans and actions? These assessments are typically made under conditions of uncertainty because relevant information may be missing or even unknowable (such as a foreign leader's intentions), information collection may be biased, and information may be unreliable as well as purposefully misleading. Therefore, most analytic assessments are essentially subjective probability judgements. Not only are analysts expected to accurately judge the probabilities, they are also expected to communicate these probabilities and their confidence in them in a veridical way to intelligence consumers. These consumers include other analysts who rely on prior intelligence reporting as well as policymakers and other decision makers (e.g., military commanders) who make critical decisions about defence and security.

The unambiguous communication of probability in intelligence assessments is important because miscommunication can lead to 'knock-on' errors in intelligence reporting that may, in turn, prompt erroneous and biased decision making. These consequences were acutely evident in the 2003 US-led coalition invasion of Iraq that aimed to find Weapons of Mass Destruction (WMD) which, in fact, did not exist. Post-mortem analyses of this major intelligence failure criticized intelligence organizations for understating the 'uncertainty' in their assessments and suggesting greater certainty than was warranted by the available information [1]; [2]; [3], p.173. The Butler report [4] further questioned whether intelligence products were "drafted and presented in a way which best helps readers to pick up the range of uncertainty attaching to intelligence assessments". The issue of knock-on errors in reporting was highlighted in the 2004 US Congressional report which stated that building on prior intelligence "without carrying forward the uncertainty from the first layer,… gave the impression of greater certainty about its judgments than was warranted" [3], p. 144. The deleterious effects of this intelligence on decision making was emphasized in the Chilcot inquiry which quoted UK Prime Minister Tony Blair's belief that intelligence organizations were "sure" about Iraq possessing WMD [5].

As the aforementioned quotes suggest, discussions of uncertainty in intelligence assessments sometimes conflate the concepts of probability and confidence. However, both analysts and intelligence organizations appear to find it difficult to grapple with these two concepts. For instance, Friedman and Zeckhauser [6] found that analysts confounded probability judgements with judgements of confidence. Similarly, the US Joint Chiefs of Staff Joint Publication (JP) 2-0 Joint Intelligence [7], which is used for joint and multinational intelligence in military operations, states that confidence should be expressed in three levels using a variety of linguistic probabilities (e.g., low confidence may be expressed by 'possible', moderate confidence may be expressed by terms such as 'likely' and 'unlikely', and high confidence may be expressed by terms such as 'almost certainly' and 'remote'). In the present chapter, we focus on the Intelligence Community's (IC's) communication of probability.

Some in the IC may believe that the problems associated with the miscommunication of probability can be resolved by simply avoiding it altogether. This can be done either by suggesting complete certainty or by neglecting to communicate probability. Kesselman [8] searched the contents of National Intelligence

Estimates (NIEs) (otherwise called intelligence reports or products) written between the 1950s and 2000s, and Mandel and Barnes [9] analyzed the language used in 2,013 Canadian strategic intelligence forecasts produced over roughly a six-year period [10]. These researchers found that 'will', which represents certainty, was the most commonly used word by analysts. Friedman and Zeckhauser [6] found that a notable proportion (i.e., 18%) of NIEs written between 1964 and 1994 did not provide any assessment of the probability associated with the outcome being forecast. However, neglecting probability is no panacea for the problems of miscommunicating it. For example, post-mortem analyses of the failed US invasion of the Bay of Pigs in 1961, which aimed to overthrow Castro's communist regime in Cuba, pointed to, amongst other things, the neglect to communicate to decision makers the chances of success without overt support from the US military in what was meant to be a clandestine operation [11], [12]. Probability should neither be avoided nor neglected, as acknowledged in the 2004 US Congressional report which states that "As much as they hate to do it, analysts must be comfortable facing up to uncertainty and being explicit about it in their assessments" [3], p. 408.

In an effort to encourage analysts to communicate the probabilities in their assessments in an unambiguous way, intelligence organizations have, in recent years, implemented formal policies for such communications. Virtually all of these policies advise analysts to use linguistic probabilities (e.g., terms such as 'very likely'). In this chapter, we critically review these policies in the US and UK – nations that share their intelligence products with one another and nations that produce joint intelligence. We point to their commitment to communicating probabilities using what are essentially vague linguistic probabilities, in spite of the well-documented pitfalls in doing so. We also point to a seeming reluctance to effectively exploit relevant scientific evidence when developing and refining methods for communicating probability, which we argue has resulted in less than optimal methods for communicating uncertainties.

17.2 EXISTING POLICIES FOR COMMUNICATING PROBABILITY IN INTELLIGENCE ANALYSIS

The US and UK introduced policies for communicating probability in intelligence assessments in the mid-2000s following recommendations made by inquiries into the Iraq WMD intelligence failure. In the UK, the 2004 Butler review stated, "While not arguing for a particular approach to [expressing]…uncertainty, …we recommend that the intelligence community review their conventions again to see if there would be advantage in refreshing them" [4]. In the US, the 2004 Congressional report stated, "Whatever device is used to signal the degree of certainty – mathematical percentages, graphic representations, or key phrases – all analysts in the Community should have a common understanding of what the indicators mean and how to use them" [3], p. 419.

Policies were developed in the US by the Office of the Director of National Intelligence (ODNI), and in the UK by Defence Intelligence (DI) originally and the Professional Head of Intelligence Analysis (PHIA) more recently. Both ODNI and PHIA are responsible for overseeing the IC in their respective countries. Despite the various options for communicating probability available to them, both the US and UK chose to convey probabilities in intelligence assessments using linguistic probabilities. Specifically, analysts in both countries (as in many others that follow a comparable approach) are required to use a standardized lexicon – namely, a list of selected terms or phrases that are ordered from the lowest to the highest degree of probability. In some cases, the phrases are combined with numeric values, which could be point estimates (e.g., 80%; see Ref. [13]), but are typically ranges (e.g., 55% – 80%). Sherman Kent, who played a key role in founding the Central Intelligence Agency's (CIA's) Office of National Estimates (the predecessor of today's US National Intelligence Council), was the first to advocate this type of standardized lexicon, although it was not formally implemented in the CIA at the time [14].

The lexicons in the US and UK have undergone several revisions over the years. Table 17-1 and Table 17-2 show the most recent versions of the US and UK lexicons [16], [17]. As can be seen in Table 17-1, the US

lexicon comprises seven categories containing a total of 14 phrases that represent numeric values from 1% to 99%. Table 17-2 shows that the UK lexicon comprises seven categories containing a total of eight phrases which are mapped onto approximate numeric ranges in two formats; namely, percentages and fractions. The two formats are only approximately synonymous. For instance, 35% is close to the fractional "equivalent" of 1/3, but not exact. However, this slight disparity is usually made salient in visual displays of the UK standard.

Table 17-1: Standardized Lexicon Developed by ODNI. Used since January 2015.

Phrase	Numeric Value (%)
Almost no chance / remote	1 – 5
Very unlikely / highly improbable	5 – 20
Unlikely / improbable (improbably)	20 – 45
Roughly even chance / roughly even odds	45 – 55
Likely / probable (probably)	55 – 80
Very likely / highly probable	80 – 95
Almost certain(ly) / nearly certain	95 – 99

Table 17-2: Standardized Lexicon Developed by PHIA. Also called the 'probability yardstick'. Used since March 2018.

Phrase	Percentage Value (%)	Fraction Value
Remote chance	$\leq \approx 5$	$\leq \approx 1/20$
Highly unlikely	$\approx 10 - \approx 20$	$\approx 1/10 - \approx 1/5$
Unlikely	$\approx 25 - \approx 35$	$\approx 1/4 - \approx 1/3$
Realistic possibility	$\approx 40 - < 50$	$\approx 4/10 - < 1/2$
Likely/probably	$\approx 55 - \approx 75$	$\approx 4/7 - \approx 3/4$
Highly likely	$\approx 80 - \approx 90$	$\approx 4/5 - \approx 9/10$
Almost certain	$\geq \approx 95$	$\geq \approx 19/20$

Note: The symbol ≈ means "approximately equal to".

Although these standards are national in breadth of applicability, they are implemented with slight variations across organizations that use the relevant standard. For instance, the National Crime Agency in the UK uses a visual representation of the yardstick in which the verbal probability terms are shown along a horizontal line ranging from 0% to 100% [17]. The boundaries of each term are stated as percentage chances but not as fractions. Probabilities below 50% are depicted in shades of blue, whereas probabilities above 50% are depicted in shades of purple.

Although we believe that the IC should develop clear guidelines for communicating aspects of uncertainty such as probabilities and confidence levels, we are concerned by the solutions they have adopted. In the remainder of this section, we summarize many of these concerns as they pertain to specific features of the current US and UK lexicons. Although the two lexicons exhibit differences, they share the same general approach, which entails selecting a circumscribed set of verbal probability terms and then attempting to set

bounds on the meaning of those terms by aligning them with numeric probability intervals. Specifically, both lexicons provide seven categories of probability from which analysts may select. This has the effect of coarsening the 0 – 1 probability scale and can artificially inflate the expressed likelihood of rare events, tail risk, or what has otherwise been described as black swans – namely, events that have extremely low probabilities but also extremely high consequence severities [18]. Indeed, while phrases such as 'remote chance' and 'almost no chance' are the lowest probability categories in the lexicons, these do not convey exceedingly small chances that would be required to accurately characterize tail risks. In fact, they are likely to be orders of magnitude off. For instance, a study of the interpretability of such terms found that the peak interpretation of 'remote chance' by a sample of intelligence analysts in the UK was about 23% and it was about 17% in a sample of Canadian analysts [19]. Clearly, an analyst could not effectively communicate a 1% chance using this term, let alone a 1/10,000th chance, which is still large in the realm of black swans. We believe IC standards for communicating probabilities to decision makers should enable analysts to be as granular in their assessment as they may need to be.

Another concern we have with communication methods that rely on verbal probabilities is that such terms do not only convey probability, they also convey "direction" to recipients. Directionality refers to whether the phase itself is positive or negative. For instance, strengthening a positive term, such as likely, for instance, by using "very" or "highly", will raise the probability conveyed, whereas doing the same to a negative term will lower the probability. This in itself is not problematic. However, research has shown that phrases with negative directionality are interpreted with greater variability than positively worded phrases [20]. This is supported by evidence from research on intelligence analysts [19], [21], [22].

All categories in the US lexicon and one category in the UK lexicon contain more than one phrase and so are intended to be fully substitutable. It has been argued that providing synonyms allows for stylistic expression in intelligence reports [13]. However, some evidence suggests that synonyms in the lexicons may not be interchangeable in the minds of analysts [10], [21], [22]. Conversely, recipients of communications might treat terms that are differentiated in the lexicons as near synonymous. Ho *et al.* [19], for instance, found that intelligence analysts treated remote chance and very unlikely as nearly synonymous. However, we see that in both the US and UK lexicons these terms occupy adjacent ordinal positions in their respective scales.

The verbal terms in both lexicons are associated with numeric ranges. This is intended to provide boundaries on the inherent fuzziness of the probability terms' meanings. However, Budescu Por, Broomell, and Smithson [23] found that recipients of verbal probability phrases in lexicons often do not keep the standards in mind and default to their personal interpretation of these phrases, thus defeating their purpose. These authors also showed that variability in the understanding of linguistic probabilities could be reduced somewhat by presenting the numeric range a phrase is intended to represent alongside the linguistic expression of probability in the statement itself. However, if numeric ranges had to be included in intelligence estimates in order to give the verbal terms clear enough meaning, then it begs the question why use the terms at all? Why not simply present the ranges, which presumably could then be specifically tailored to the assessment? In addition, there are currently some categories that have sizeable numeric ranges (i.e., up to 25% points in the US lexicon and 20% points in the UK lexicon). Consequently, some phrases are much fuzzier than others. Wide latitude for interpretation increases opportunities for misunderstanding of probability. For instance, whereas a US analyst writing an intelligence report may intend 'likely' to mean 55% (i.e., bottom of the range), a reader of his/her report may understand it to mean 80% (top of the range).

The current standards, while adopting the same general approach, differ in terms of many specific aspects that can undermine bilateral interoperability and cause communication errors. For instance, the US lexicon contains 14 phrases organized into two sets, whereas the UK lexicon has a single set of eight phrases. Only five phrases are shared between the lexicons and, notably, the UK lexicon contains the phrase 'realistic possibility'. Barnes [13] "banned" the use of the phrase 'realistic possibility' in his Canadian probability lexicon because of his belief that it lacked precision (see also Kent's 1964 discussion of weasel words [14]). Indeed, studies reveal that US and UK analysts do not normally use this phrase to express probability [21], [22].

Finally, the US lexicon does not cover the end-points of the probability interval (i.e., 0% and 1%), and whereas the UK lexicon does, it has gaps between categories (e.g., 6% – 9%). Complete certainty therefore, cannot be expressed linguistically using the US lexicon, and specific points along the probability scale cannot be expressed using the UK lexicon. Similarly, omission of the end-points precludes all orders of magnitude between .99 and 1, and .01 and 0. In the US lexicon, each category slightly overlaps the next, whereas in the UK lexicon, as mentioned, there are gaps between categories. Analysts using the US lexicon may find it difficult to select specific phrases when communicating probability at the top or bottom ends of a category. This is particularly problematic given the wide category ranges. For example, an analyst wishing to express a numeric value of 55% could use 'roughly even odds' which has a bottom range of 45% or 'probable,' which has a top range of 80%. Clearly, a decision maker may respond differently if he/she believes the likelihood being expressed is 45% or 55% versus 80%. Finally, whereas categories in the US lexicon are associated with numeric ranges, in the UK lexicon they are associated with both numeric and fraction ranges. The denominators in these fractions differ both within and across categories (e.g., the second category represents ≈ 1/10 – ≈ 1/5 and the third category represents ≈ 1/4 – ≈ 1/3). There is a total of seven different denominators used in the UK lexicon. The lack of a common denominator means that information cannot be easily aggregated (summed) and effort has to be made to find a common denominator.

17.3 THE IC'S COMMITMENT TO LINGUISTIC PROBABILITIES

Intelligence organizations strongly prefer to communicate probability linguistically rather than numerically. In their analysis of NIEs, Friedman and Zeckhauser found that only four percent contained numeric expressions of probability (e.g., percentages, odds) [6]. Thus, the lexicons introduced in the UK and US in mid-2000s merely institutionalize the IC's longstanding bias against precision and quantification of probability. The IC is not alone in the preference for using linguistic probabilities. Indeed, research has demonstrated that people generally prefer to communicate probability linguistically rather than numerically [24], [25], [26]. This is particularly so when judgements are made under conditions characteristic of the intelligence analysis domain, namely, where people sense that relevant information is unreliable or unknown [27].

People may prefer to express probability linguistically rather than numerically for several reasons. Wallsten *et al.* found that some people said it was easier and more natural to use language rather than numbers [25]. For some, this preference was particularly so when the issue was deemed to be unimportant and/or the information was unreliable. In the IC, Barnes observed that analysts initially perceived it to be too difficult to estimate a numeric value [13]. Analysts were also concerned that numeric estimates would mislead by providing a false sense of precision, which might cause policymakers to put too much confidence in intelligence assessments. Finally, analysts were anxious about numeric values being used to evaluate their analytic performance because these can be easily tested (e.g., by the use of Brier scores [28]). The sentiments of individual analysts are shared by organizations overseeing the IC. For instance, ODNI stated, "Assigning precise numerical ratings to such [analytic] judgments would imply more rigor than we intend." [29].

However, the benefit of communicating probabilities with words is questionable given the significant, well-documented pitfalls of doing so. Research has repeatedly demonstrated considerable variability in how people understand linguistic probabilities in both laboratory settings and applied domains [30], [31], [32], [33], [25], [34], [35], [36]. This variability is evident across people as well as within an individual person. Specifically, individuals may have broad or fuzzy interpretations of particular phrases. In addition, different people may use different phrases to refer to the same probability value(s) and/or may use the same phrase to refer to different values.

Research also shows that the interpretation of linguistic probabilities may be affected by the context in which they are used [10], [20], [37], [38], [39], [40], [41], [42], [43]. These contexts can be externally provided (e.g., the base rate of an event or the severity of an outcome) or internal to the person (e.g., one's attitude to the subject matter). In the intelligence analysis context, Mandel [10] reported that a combined sample of

intelligence analysts and university students had significantly less discriminating numeric interpretations of linguistic probabilities when they were used to describe an action that was described as failing than when the same action was described as succeeding.

Finally, it is unclear how effectively linguistic probabilities can be combined. This is of particular concern in the IC where analysts are often required to express how the probability of a number of (independent as well as interacting) events may combine to lead to an outcome or set of outcomes. Imagine, for example, a chain of four independent events that must occur in order for a particular threat scenario to manifest. In one case, the probabilities of the events are estimated to be .75, .10, .70, and .01. In another case, they are given the US standard equivalents, 'likely', 'very unlikely', 'probable', and 'almost no chance'. If asked, what the probability of the four events occurring was, it would be easy for the analyst using numeric probabilities to calculate the probability of the conjunction, which is simply the product of the four terms, equalling 5¼ out of 10,000. We strongly suspect that the analyst using the linguistic probabilities instead would be led astray, not only because the verbal terms cannot convey such a small probability but also because words do not lend themselves to arithmetic operations.

The similar approaches to communicating probability adopted in the UK and US intelligence communities suggests that the IC believes that the pitfalls of using linguistic probabilities can be mitigated by adopting standardized lexicons that attempt to establish the intended meaning of the selected phrases. However, research shows that people find it difficult to suppress their normal meanings of linguistic probabilities even when lexical standards are provided for them to consult [23], [38]. In the IC, the ineffectiveness of this approach is no doubt further compounded by the use of multiple lexicons currently in use in organizations that share intelligence (e.g., the UK and US). These situations require analysts (and other intelligence consumers) to rapidly shift their use and understanding of phrases and to mentally juggle different lexicons. If the research indicates that even a single lexicon is not adhered to well, what is the likelihood that individuals receiving intelligence will make these additional adjustments? A glance at Table 17-1 and Table 17-2 illustrates the juggling act required due to organizational differences in lexicons. For instance, whereas the UK lexicon uses 'unlikely' to represent 25% – 35%, the US lexicon uses it to represent a broader range of probabilities (i.e., 20% – 45%), and so consumers of intelligence products from both countries would need to speculate whether these products intend to communicate the same or different probability. To illustrate the mental juggling act required due to within-organizational revisions of lexicons, we can refer to the UK where DI's version of the yardstick was replaced by PHIA's version on March 5th 2018. Analysts and their customers would therefore be expected to update their understanding of nearly all of the phrases in the previous lexicon, and to use both lexicons when referring to reports produced before March 5th and those produced after this date.

In short, the IC wants to allow analysts to use linguistic terms for conveying probabilities that are familiar to them, but it does not want to allow analysts to determine the meaning of those terms. In so doing, we believe the IC has overestimated its ability to institutionally mandate meaning, change it at will through successive revisions, and to achieve corresponding compliance from its analytic partners and collaborators.

17.4 THE IC SHOULD DEVELOP EVIDENCE-BASED POLICIES

Policy development in the IC has historically occurred without consideration of relevant evidential bases beyond the anecdotal 'lessons learned' following intelligence failures [44], [45]. This is also largely true of the development and refinement of policies for communicating probability in intelligence assessments. While we cannot rule out that ODNI and PHIA considered relevant research findings in the process of formulating their policies, the incongruity between where we believe the scientific findings point and the IC's adopted solutions indicates that the exploitation of relevant research findings was ineffective. We also see no evidence that the adopted methods were tested for their effectiveness prior or subsequent to their full adoption. Thus, neither ODNI nor PHIA knows if their lexicons are effective methods for communicating probability.

One way to assess the effectiveness of a standardized lexicon is to compare similarities and differences between what the lexicon advocates and what analysts do. For instance, do analysts use phrases in the lexicons as intended? Do they consider phrases to be substitutable in the same way as advocated in the lexicons? Fortunately, there is a growing body of research on how probability is communicated in the IC. Some studies are based on a content analysis of actual intelligence reports [6], [8], [9]. While the external validity of this method is high, the method is limited because it does not enable researchers to determine what probability value(s) analysts had in mind when they used a specific phrase, unless they also provided numeric values in the reports (which may be relatively uncommon given the preference for communicating probability linguistically rather than numerically).

Partly in order to overcome this limitation, other studies have used quantitative methods [10], [19], [21], [22], that elicit people's lexicons for communicating probability and measure how people numerically interpret linguistic probabilities (for details of some specific methods used, see Refs. [33], [46], [47]). Regardless of the method used, however, the extant body of research points to discrepancies between policy (what the lexicons advocate) and practice (what analysts do). Below, we summarize the main findings from this research and their implications for the US and UK lexicons.

Kesselman counted the occurrence of a pre-selected list of 63 words and phrases, including those in the original lexicon produced by ODNI in the contents of NIEs written from the 1950s to mid-2000s [8]. Her analysis revealed trends in the language used across the decades. The use of 'even chance' waned and then increased from the 1950s to 1990s. Words such as 'probably' and 'likely' that were intended to be interchangeable in the ODNI's lexicon differed in their popularity; i.e., the use of 'likely' increased from the 1950s to 1990s, whereas the use of 'probably' decreased. In fact, phrases in the ODNI's lexicon were not equally popular, and in the early to mid-2000s, there was no occurrence of 'remote' and 'even chance' in the NIEs. Although Kesselman was unable to discern the probabilities that analysts were referring to when using specific phrases, it is clear that the ODNI's efforts to reinstitute language that had naturally gone 'out of fashion' may face challenges to adoption.

Mandel [10] examined a standardized lexicon used by strategic analysts in a Canadian strategic intelligence unit, which for present purposes included some of the phrases in the ODNI's and PHIA's current lexicons. He asked a combined sample of military intelligence analysts (who did not use the lexicon on the job) and university students to provide the best numeric probability equivalent for each of the phrases. His findings indicate that interpretations of the phrases in the current ODNI and PHIA lexicons did not map directly onto the prescribed numeric ranges. In some cases, credible ranges based on the 95% confidence interval around participants' median equivalency estimates were either below or above the category range. For instance, in PHIA's lexicon, whereas 'highly unlikely' represents 10% – 20%, the corresponding 95% confidence interval on the median best estimate in Mandel's study was 8% – 10%. Similarly, whereas 'likely' is intended to represent 55% – 75% in PHIA's lexicon, the 95% confidence interval on the median estimate of this term was 75% – 80%. Finally, phrases such as 'likely' and 'probably/probable' that the lexicons intend to be substitutable were not perfectly so in participants' minds [10]. Thus, Canadian analysts relying on reports produced by the US and UK may misunderstand the probabilities that their American and British counterparts intend to communicate. Of course, misunderstanding can also occur amongst other consumers of US and UK intelligence products.

Wallsten *et al.* [22] examined the linguistic probability lexicons of 119 CIA analysts. They found that the average size of the lexicon (spanning the 0 to 1 probability interval) was approximately eight phrases. This is much less than the current ODNI lexicon and equivalent to the current PHIA lexicon. Wallsten *et al.* also found considerable variability in the phrases that appeared in analysts' lexicons (i.e., there were 170 distinct phrases). 'Unlikely' and 'likely' were the most common phrases, appearing in around 70 analysts' lexicons. For present purposes, we point to the fact that although there was a general correspondence between the rank order of phrases in analysts' lexicons and ODNI's and PHIA's prescribed rank order, analysts' numeric interpretations of phrases in these lexicons again differed from that intended, with some phrases being

interpreted outside the category ranges. The fact that US analysts may not interpret phrases as prescribed by ODNI suggests that expecting them to suppress their vernacular interpretation of linguistic probabilities may be unrealistic.

Most recently, Dhami [21] found comparable results in a study of the linguistic probability lexicons of 26 UK intelligence analysts who are responsible for writing intelligence reports. Analysts' lexicons contained, on average, ten phrases, and a total of 145 unique phrases were used. 'Likely' and 'unlikely' were, again, some of the most commonly used phrases. In addition, although the rank order of phrases in analysts' lexicons generally corresponded to the rank order in the US and UK lexicons, analysts whose lexicons contained phrases that appeared in these lexicons did not use them as mandated. For example, while 'likely' was ranked before 'highly unlikely' in analysts' lexicons, they used the phrases to represent ranges with higher maximum probabilities than that mandated in the UK lexicon. Interpretations of some phrases were above or below the category ranges prescribed in PHIA's current lexicon, and analysts did not consider 'likely' and 'probably' to be interchangeable as intended.

Finally, Ho *et al.* [19] studied 61 Canadian and UK intelligence analysts' interpretations of phrases in the US and UK lexicons, in an effort to compare the effectiveness of an evidence-based lexicon versus the US and UK lexicons. The researchers used one group of analysts to derive different evidence-based lexicons that utilized the same probability phrases as those common to the US and UK lexicons. The evidence-based lexicons relied on different statistical methods to set the numeric ranges that corresponded to each probability phrase. Ho *et al.* then studied a distinct group of analysts and recorded the proportion of their numeric probability equivalency estimates that fell in the ranges stipulated by the various lexicons. The UK and best performing evidence-based lexicon outperformed the US lexicon at the extremes (i.e., for the lowest and highest probability phrases), whereas the US and evidence-based lexicon outperformed the UK lexicon for the mid-range probabilities. Overall, Ho *et al.* [19] showed that the evidence-based lexicon was superior to both the UK and US lexicons in capturing analysts' interpretations of the institutionally mandated terms.

17.5 CONCLUSION AND THE WAY FORWARD

In the present chapter, we highlighted a number of significant pitfalls associated with using linguistic probabilities to communicate probability. These are due to the variability in the interpretation of phrases, the effect of contextual factors on interpretation of phrases, and the difficulties of aggregating the meaning of phrases. Requiring analysts to suppress their normal meanings of phrases when using standardized lexicons such as those advocated by ODNI and PHIA is deeply problematic and virtually untenable. An alternative approach to communicating probability is warranted that draws on scientific theory and evidence and which would overcome the pitfalls associated with using linguistic probabilities. The fact that the miscommunication of probability in intelligence products can have devastating consequences underscores the need to improve current policies and practices.

Perhaps the default policy ought to be to require analysts to use their own selection of numeric ranges (and precise numeric probabilities where appropriate). Intelligence organizations should not only encourage and permit analysts to be more precise and clearer in their communication of probabilities, but they should also incentivize analysts to perform analysis in a way that helps them to generate more precise probabilities. Recall that ODNI eschewed the use of numeric probabilities in the US because they believed this may imply greater rigor than intended [29]. We argue that analysis ought to be rigorous even when the bases for assessment are subjective probability judgements. It remains a deep misconception in the IC that only frequentistic probabilities (or relative frequencies) can be quantified. This flies in the face of Bayesianism and the very foundations of rational choice theory [48]. We recommend that, at minimum, the IC invest in research that examines the relative effectiveness of current methods and alternatives that involve the use of numeric probabilities. We believe that the results of such a relatively low-cost research program would be highly useful.

17.6 REFERENCES

[1] UK House of Commons Foreign Affairs Committee. (2003). *The Decision to Go to War in Iraq.* Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/272087/6062.pdf

[2] UK Intelligence and Security Committee. (2003). *Iraqi Weapons of Mass Destruction – Intelligence and Assessments.* Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/272079/5972.pdf

[3] US Congressional Select Committee on Intelligence. (2004). *Report on the U.S. Intelligence Community's Prewar Intelligence Assessments on Iraq.* Washington DC Retrieved from https://fas.org/irp/congress/2004_rpt/ssci_iraq.pdf

[4] The Lord Butler of Brockwell, Chilcot, J., The Lord Inge, Mates, M., and Taylor, A. (2004). *Review of Intelligence on Weapons of Mass Destruction: Implementation of Its Conclusions*. Retrieved from https://fas.org/irp/world/uk/butler071404.pdf

[5] Chilcot, J. (2016). *The Report of the Iraq Inquiry. Executive Summary.* Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/535407/The_Report_of_the_Iraq_Inquiry_-_Executive_Summary.pdf

[6] Friedman, J.A., and Zeckhauser, R. (2012). Assessing uncertainty in intelligence. *Intelligence and National Security,* 27(6):824-847.

[7] United States Joint Chiefs of Staff. (2013). Joint Publication JP 2-0: *Joint Intelligence,* Washington, DC. Retrieved from https://fas.org/irp/doddir/dod/jp2_0.pdf.

[8] Kesselman, R.F. (2008). Verbal probability expressions in national intelligence estimates: A comprehensive analysis of trends from the fifties to post 9/11. Master's thesis. Erie, PA: Mercyhurst College.

[9] Mandel, D.R., and Barnes, A. (2018). Geopolitical forecasting skill in strategic intelligence. *Journal of Behavioral Decision Making,* 31(1):127-137.

[10] Mandel, D.R. (2015). Accuracy of intelligence forecasts from the intelligence consumer's perspective. *Policy Insights from the Behavioral and Brain Sciences,* 2(1):111-120.

[11] Pfeiffer, J.B. (1984). *The Taylor Committee Investigation of the Bay of Pigs.* Retrieved from https://nsarchive2.gwu.edu/NSAEBB/NSAEBB355/bop-vol4.pdf.

[12] US Inspector General. (1962). *Inspector General's Survey of the Cuban Operation October 1961.* Retrieved from https://nsarchive2.gwu.edu//NSAEBB/NSAEBB341/IGrpt1.pdf.

[13] Barnes, A. (2016). Making intelligence analysis more intelligent: Using numeric probabilities. *Intelligence and National Security,* 31(3):327-344.

[14] Kent, S. (1964). Words of estimative probability. *Studies in Intelligence*, 8(4):49-65.

[15] UK Defence Intelligence. (n.d.). *PHIA Probability Yardstick.* London, UK. (n.p.).

[16] Office of the Director of National Intelligence. (2015). *Intelligence Community Directive ICD 203, Analytic Standards*. Retrieved from https://fas.org/irp/dni/icd/icd-203.pdf.

[17] National Crime Agency. (2018). *National Strategic Assessment of Serious and Organized Crime 2018*. Retrieved from http://www.nationalcrimeagency.gov.uk/publications/905-national-strategic-assessment-for-soc-2018/file.

[18] Makridakis, N., and Taleb, N. (2009). Living in a world of low levels of predictability. *International Journal of Forecasting,* 25(4):840-844.

[19] Ho, E., Budescu, D.V., Dhami, M.K., and Mandel, D.R. (2015). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science & Policy,* 1(2):43-55.

[20] Smithson, M., Budescu, D.V., Broomell, S.B., and Por, H. (2012). Never say "not": Impact of negative wording in probability phrases on imprecise probability judgments. *International Journal of Approximate Reasoning,* 53(8):1262-1270.

[21] Dhami, M.K. (2018). Towards an evidence-based approach to communicating uncertainty in intelligence analysis. *Intelligence and National Security,* 33(2):257-272.

[22] Wallsten, T.S., Shlomi, Y., and Ting, H. (2008). *Exploring Intelligence Analysts' Selection and Interpretation of Probability Terms: Final Report for Research Contract 'Expressing Probability in Intelligence Analysis'*. Sponsored by the CIA.

[23] Budescu, D.V., Por, H., Broomell, S.B., and Smithson, M. (2014). Interpretation of IPCC probabilistic statements around the world. *Nature Climate Change,* 4(6):508-512.

[24] Brun, W., and Teigen, K. (1988). Verbal probabilities: Ambiguous, context dependent, or both? *Organizational Behavior and Human Decision Processes,* 41(3):390-404.

[25] Erev, I., and Cohen, B.L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes,* 45(1):1-18.

[26] Wallsten, T.S., Budescu, D.V., Zwick, R., and Kemp, S.M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society,* 31(2):135-138.

[27] Olson, M.J., and Budescu, D.V. (1997). Patterns of preference for numerical and verbal probabilities. *Journal of Behavioral Decision Making,* 10(2):117-131.

[28] Mandel, D.R., and Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences of the United States of America,* 111(30):10984-10989.

[29] Office of the Director of National Intelligence. (2007). *National Intelligence Estimate – Prospects for Iraq's Stability: A Challenging Road Ahead*. National Intelligence Estimate. Retrieved from https://www.dni.gov/files/documents/Newsroom/Press%20Releases/2007%20Press%20Releases/20070202_release.pdf

[30] Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting,* 1(3):257-269.

[31] Budescu, D.V., Weinberg, S., and Wallsten, T.S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance* 14:281-294.

[32] Clarke, V.A., Ruffin, C.L., Hill, D.J., and Beaman, A.L. (1992). Ratings of orally presented verbal expressions of probability by a heterogenous sample. *Journal of Applied Social Psychology,* 22(8):638-656.

[33] Dhami, M.K., and Wallsten, T.S. (2005). Interpersonal comparison of subjective probabilities. *Memory & Cognition,* 33(6):1057-1068.

[34] Karelitz, T., and Budescu, D.V. (2004). You say "probable" and I say "likely": Improving interpersonal communication with verbal probability phrases. *Journal of Experimental Psychology: Applied,* 10(1):25-41.

[35] Reagan, R.T., Mosteller, F., and Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of Applied Psychology,* 74(3):433-442.

[36] Zwick, R., and Wallsten, T.S. (1989). Combining stochastic uncertainty and linguistic inexactness: Theory and experimental evaluation of four fuzzy probability models. *International Journal of Man-Machine Studies,* 30(1):69-111.

[37] Fischer, K., and Jungerman, H. (1996). Rarely occurring headaches and rarely occurring blindness: Is rarely=rarely? The meaning of verbal frequentistic labels in specific medical contexts. *Journal of Behavioral Decision Making,* 9(3):153-172.

[38] Budescu, D.V., Por, H., and Broomell, S.B. (2012). Effective communication of uncertainty in IPCC reports. *Climate Change,* 113(2):181-200.

[39] Fox, C.R., and Irwin, J.R. (1998). The role of context in the communication of uncertain beliefs. *Basic and Applied Social Psychology,* 20(1):57-70.

[40] Harris, A., and Corner, A. (2011). Communicating environmental risks: Clarifying the severity effect in interpretations of verbal probability expressions. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 37(6):1571-1578.

[41] Piercey, M.D. (2009). Motivated reasoning and verbal vs. numerical probability assessment: Evidence from an accounting context. *Organizational Behavior and Human Decision Processes,* 108(2):330-341.

[42] Wallsten, T.S., Fillenbaum, S., and Cox, A. (1986). Base-rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language,* 25(5):571-587.

[43] Weber, E.U., and Hilton, D.J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance,* 16(4):781-789.

[44] Dhami, M.K., Mandel, D.R., Mellers, B.A., and Tetlock, P.E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science,* 10(6):753-757.

[45] Mandel, D.R., and Tetlock, P.E. (2018). Correcting judgment correctives in national security intelligence. *Frontiers in Psychology* 9:2640.

[46] Karelitz, T., Budescu, D.V., and Wallsten, T.S. (2000). *Validation of a new technique for eliciting membership functions of probability phrases*. Poster presented at the Meeting of the Society for Judgment and Decision Making, New Orleans, LA.

[47] Wallsten, T.S. (1971). Subjectively expected utility theory and subjects' probability estimates: Use of measurement-free techniques. *Journal of Experimental Psychology,* 88(1):31-40.

[48] Savage, L.J. (1954). *The Foundations of Statistics.* New York, NY: Wiley.

Chapter 18 – VARIANTS OF VAGUE VERBIAGE: INTELLIGENCE COMMUNITY METHODS FOR COMMUNICATING PROBABILITY[1,2]

Daniel Irwin and David R. Mandel
Defence Research and Development Canada
CANADA

18.1 INTRODUCTION

Intelligence analysts are often tasked with estimating the probability of a future development (e.g., candidate *x* is likely to win the election), or that an explanation for a past/ongoing development is true (e.g., it is unlikely that country *y* armed group *z*) [2]. In order to provide meaningful decision support, these estimates must not only be derived from accurate information and sound reasoning, they must also be communicated effectively to decision makers [3], [4]. In command and national security decision making, imprecise or ambiguous estimates may precipitate intelligence failure. For instance, in 1961, the US Joint Chiefs of Staff evaluated plans for a CIA-backed operation to overthrow Fidel Castro. In their report, the Joint Chiefs wrote that the attack had a "fair chance" of success, despite agreeing internally on 3:10 odds [5]. While the authors of the report believed "fair chance" would be interpreted as "not too good," President John F. Kennedy viewed the assessment as favorable, and proceeded to authorize the disastrous Bay of Pigs invasion [5].

Debates over uncertainty communication have long pervaded the intelligence discourse (e.g., Refs. [6], [7], [8]) and became acute following major intelligence failures; namely, the 9/11 attacks and Iraq WMD fiasco [9], [10]. To address the goal of mitigating subjectivity and the potential for miscommunication, some intelligence organizations have developed standardized lexicons for communicating estimative probability. However, as is the case with standards developed for other facets of uncertainty communication in intelligence (e.g., information credibility, source reliability, analytic confidence), these standards are rarely grounded in empirical research, and they may in fact undermine rather than improve communication fidelity [4], [10], [11]. One overriding reason for our pessimistic outlook is that, for the most part, the set of current standards represents various versions of vague verbiage: all current intelligence standards for communicating estimative probabilities commit to using verbal probabilities (words such as "likely" or phrases like "realistic possibility") and they shun the use of numerical probabilities either as precise estimates (e.g., 73% chance) or imprecise estimates (e.g., a 60% – 80% chance).

In this chapter, we present an annotated collection of the estimative probability standards gathered by members and affiliates of NATO's SAS-114 Research Task Group on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making. These include standards used in intelligence production, as well as in other domains such as defence and security risk management and climate science. After reviewing this non-exhaustive collection of standards, we discuss common problematic features and how they might compromise efforts to support decision making (see also chapters by Dhami and Mandel and by Rein in this report).

[1] This chapter expands on the following client-focused document Methods for Communicating Estimative Probability in Intelligence to Decision Makers: An Annotated Collection [1].

[2] Funding support for this work is provided by the Canadian Safety and Security Program Project CSSP-2016-TI-2224 (Improving Intelligence Assessment Processes with Decision Science).

18.2 OVERVIEW OF CURRENT STANDARDS

18.2.1 National Security Intelligence Standards

18.2.1.1 NATO Standards

NATO Allied Joint Doctrine for Intelligence Procedures (NATO AJP-2.1) outlines uncertainty communication procedures intended for use by NATO members as well as external partners [12]. Estimative probability is expressed using five verbal terms with associated probability ranges. These ranges overlap at adjacent boundaries (and only at the boundaries) and they vary in size. The terms are meant to cover the full range of the probability interval [0, 1]. This is one of the coarsest probability lexicons examined, as many have seven or more terms. Analysts are explicitly discouraged from ever using the term *confirmed* (which is omitted from the standard), "given the nature of intelligence projecting forward in time." See Table 18-1

Table 18-1: NATO AJP-2.1 2016 Probability Levels [12].

Probability Statements for Assessments (Numerical and Verbal)	
More than 90%	Highly likely
60% – 90%	Likely
40% – 60%	Even chance
10% – 40%	Unlikely
Less than 10%	Highly unlikely

18.2.1.2 Canadian Standards

The former standard employed by the Middle East and Africa Division of the Privy Council Office Intelligence Assessment Secretariat (henceforth abbreviated IAS MEA) rates estimative probability on a nine-level scale, with associated qualitative terms (Table 18-2) [2]. Numerical values assigned by analysts do not appear in finished intelligence products, and are used for internal review purposes only, such as monitoring the accuracy of forecasts [13], [14]. The inclusion of synonyms is intended to provide stylistic flexibility in presenting analysis. The scale includes *will* / *is certain* and *will not* / *no prospect* to convey judgements that are deemed to be certain or very near certain. Mandel [4] found that the average interpretation of the verbal probability terms used in this standard (operationalized as the median best numerical probability equivalent given for a term) was well matched to the numerical interpretations of those terms stipulated in the standard. The high degree of agreement is likely due to the fact that Barnes [2] based those values on a careful review of the scientific literature examining how such terms are translated to numerical probabilities or probability ranges [15].

Table 18-2: IAS MEA Probability Mapping Standard [2].

Verbal Expression	Probability	Remarks
Will Is certain	[10/10]	There is no plausible scenario – however remote – where this event would not happen.
Almost certain Extremely likely Highly likely	[9/10]	There remains some conceivable scenario – albeit very remote – that this event would not happen.

Verbal Expression	Probability	Remarks
Likely Probable, probably	[7/10 – 8/10]	
Slightly greater than even chance	[6/10]	Use rarely, only when there is a specific reason to judge the probability at greater than even but cannot be categorized as 'likely'.
Even chance	[5/10]	
Slightly less than even chance	[4/10]	Use rarely, only when there is a specific reason to judge the probability at less than even but cannot be categorized as 'unlikely'.
Unlikely (Only a) low probability Probably not	[2/10 – 3/10]	
Very unlikely Highly unlikely Extremely unlikely Little prospect	[1/10]	There remains some conceivable scenario – albeit very remote – that this event could happen.
No prospect Will not	[0/10]	There is no plausible scenario – however remote – where this event could happen.

In the Canadian Forces Intelligence Command (CFINTCOM) Aide-Mémoire on Intelligence, estimative probability is gauged on an eleven-point scale with associated qualitative terms (Figure 18-1) [16]. These values "are not intended to express a percentage or other numerical interpretation but are simply to indicate likelihood" [16]. The scale is based on an earlier version of the IAS MEA scale, but it also incorporates verbal probability expressions from Defense Intelligence Agency (DIA) Tradecraft Note 01-15 [17]. These expressions were not suggested by Barnes [2] and lack empirical support, so their inclusion may undermine the effectiveness of the scale. The document emphasizes the separation of estimative probability and analytic confidence, but also recommends the use of specific estimative terms when conducting "exploratory analysis" (i.e., analysis with low analytic confidence) [16]. Under these circumstances, analysts are encouraged to use *possible*, *may*, *might*, *could*, *can*, *perhaps*, *unsure*, *unknown*, *do not know*, *unable to assess*, and *undetermined*. The document does not offer guidelines on mixing and matching various synonyms.

18.2.1.3 US Standards

The 2007 National Intelligence Estimate (NIE 2007) Iran: Nuclear Intentions and Capabilities [18] links probability and analytic confidence in its "What We Mean When We Say" explanation of estimative language. *Might* and *may* are highlighted as terms of estimative probability used to "reflect situations in which [analysts] are unable to assess the likelihood, generally because relevant information is unavailable, sketchy, or fragmented" (i.e., analysis with low analytic confidence). Estimative probability is also connected to the potential consequences of an event; *we cannot dismiss*, *we cannot rule out*, and *we cannot discount* are used to "reflect an unlikely, improbable, or remote event whose consequences are such that it warrants mentioning." Under these circumstances, it is unclear why analysts should not simply use the expressions *unlikely*, *improbable* or *remote*, which are already built into the lexicon (Figure 18-2). NIE 2007 forgoes the use of numerical values, instead placing qualitative terms along a spectrum. Earlier versions of this standard note that "assigning precise numerical ratings to such judgments would imply more rigor than we intended" [7].

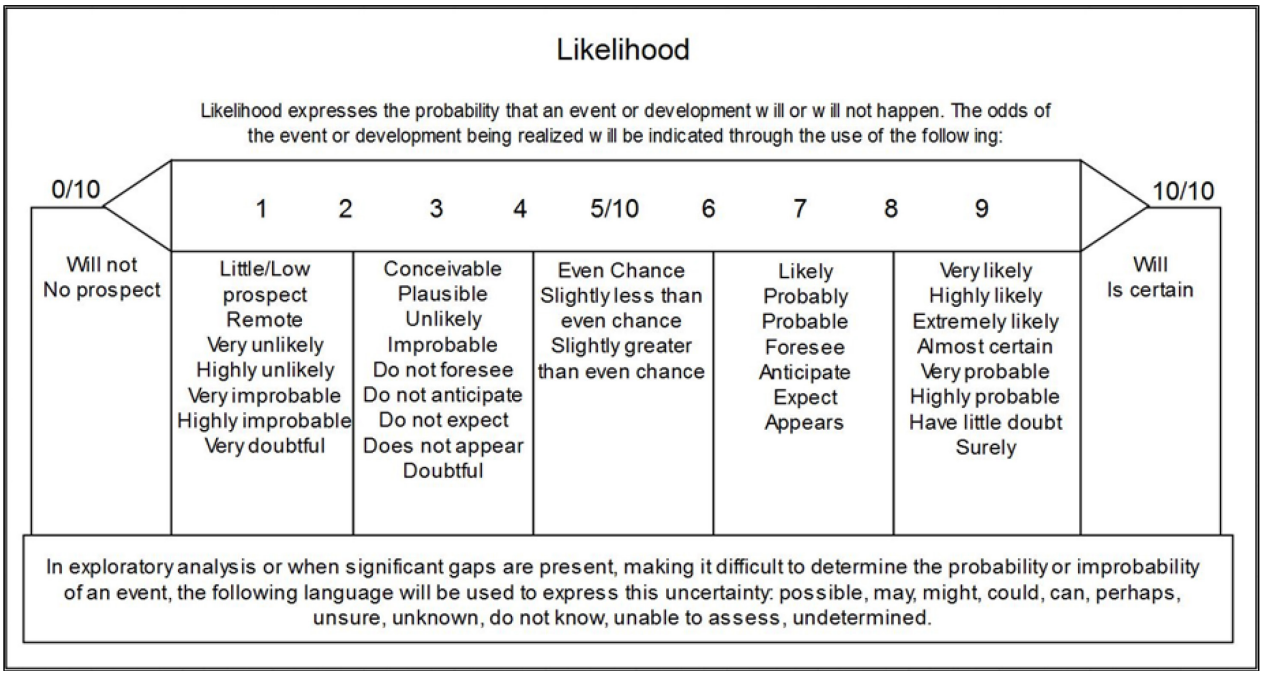


Figure 18-1: CFINTCOM Likelihood [16].

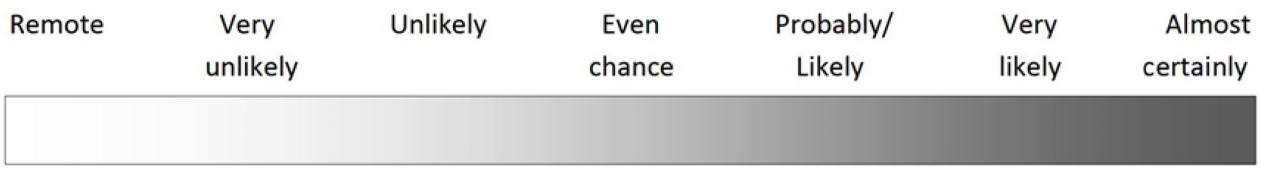


Figure 18-2: NIE 2007 Estimates of Likelihood [18].

DIA Tradecraft Note 01-15 supersedes DIA What We Mean When We Say [17]. The document cites a 13 June 2013 ODNI memo, which states that the probability of an event must be distinguished from analytic confidence. DIA conceptualizes likelihood and confidence as separate components of "Analytic Certainty." Despite this distinction, the document proscribes the use of *unlikely*, *likely*, *improbable*, and *probably* to communicate estimative probability when analytic confidence is low. The "gray zone" contains terminology for communicating an inability to gauge likelihood (Table 18-3). The document also presents strategies typically linked to analytic confidence (e.g., increased inter-analyst collaboration) that it suggests can increase or decrease the estimated probability of an event. When analysts are unable to 'move' their estimate out of the gray zone with additional information or analysis, they are required to label it "exploratory." DIA explicitly proscribes the use of percentages to communicate estimative probability. The "Expressing Analytic Certainty" graphic (Figure 18-3) is embedded in DIA products for consumer reference.

US Intelligence Community Directive (ICD) 203 provides two sets of verbal uncertainty expressions, along with numerical equivalents, which are intended for application across the US Intelligence Community [19]. The probability ranges overlap at the boundaries and vary in size. Unlike several standards examined, the numerical values in ICD 203 range from 0.01 to 0.99, rather than 0 to 1. This design may be intended to reflect the uncertain nature of intelligence estimates. ICD 203 is also the only standard examined with clear guidelines regarding the use of synonyms; analysts are explicitly discouraged from mixing terms from different rows, unless they provide a disclaimer noting that the terms indicate the same assessment of probability. ICD 203 also recommends that, to minimize confusion, analysts refrain from expressing analytic confidence and estimative probability in the same sentence.

Table 18-3: DIA Expressing Likelihood [17].

Likelihood Zones	Associated Terms
FACT	Will, definitely, sure, no doubt, positive, confirmed.
Increasing Likelihood	Very likely, highly likely, surely, very probable, highly probable, have little doubt.
	Likely, probably, foresee, see, expect, appears, anticipate (**do not select "likely" or "probably" with low confidence determinations – select different term**).
Gray Zone	Significant gaps are present making it difficult to determine the probability or improbability of an event. Identifying alternatives in grey analysis is essential. **The gray zone is exploratory analysis.** Strategies for moving out of the gray zone include: • Examining key assumptions or knowledge base for a likelihood direction; • Explain possible paths and describe indicators in each path; • Review AND update collection and knowledge base; • Executing a new or different analytic approach or methodology to gain additional insight; and • Additional collaboration. Gray words in this zone include: **possible, may, might, could, can, perhaps, unsure, we do not know, unknown, undetermined, or unable to assess**.
Decreasing Likelihood	Unlikely, improbable, do not expect, do not anticipate, do not foresee, do not see, does not appear, is doubtful (**do not select "unlikely" or "improbable" with low confidence determinations – select different term**).
	Very unlikely, highly improbable, highly unlikely, very doubtful.
FACT	Will not, definitely not, positively not; impossible.

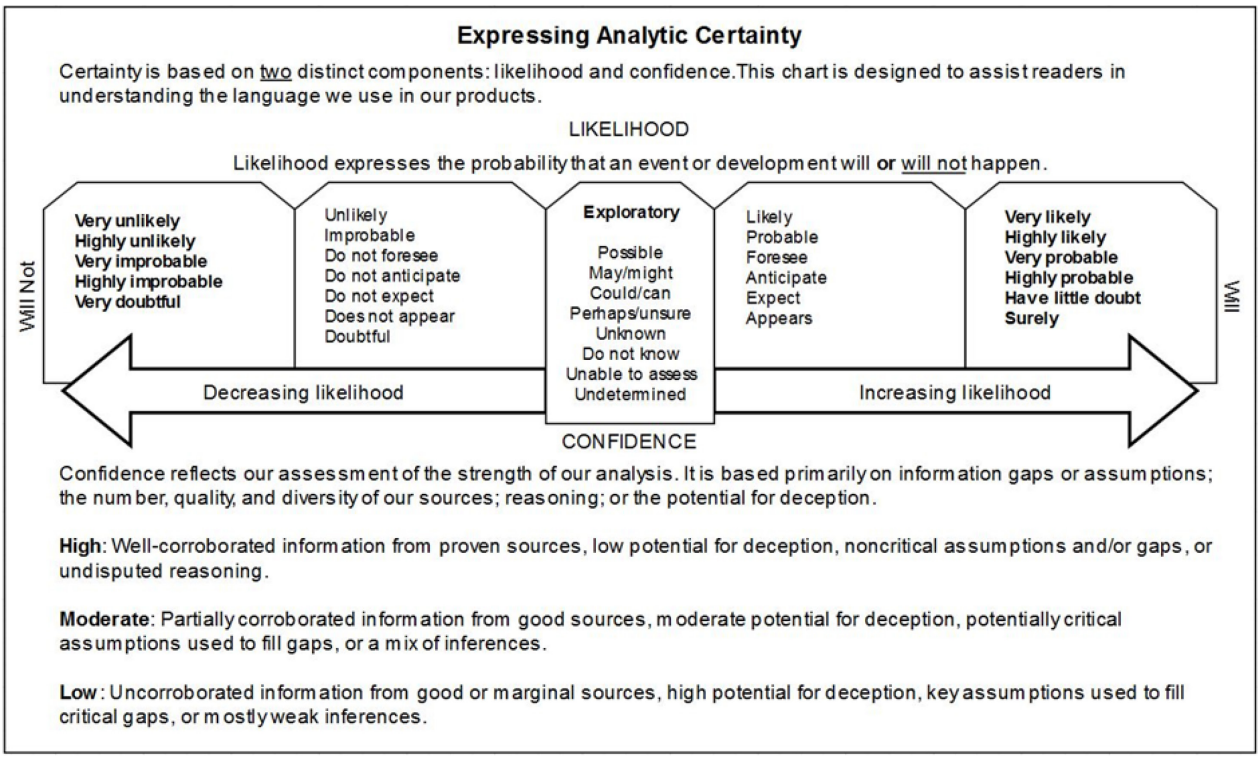


Figure 18-3: DIA Expressing Analytic Certainty [17].

National Intelligence Council (NIC) Judgments of Likelihood appears in Annex B of Intelligence Community Assessment 2017-01D [20], and illustrates how the probability scales stipulated in ICD 203 (Table 18-4) [19] are presented to US intelligence consumers (Figure 18-4). The two sets of verbal expressions outlined in ICD 203 are arranged along a coloured spectrum. The numerical probability range equivalents provided for the expressions are not shown on the spectrum representation. Moreover, the location of verbal expressions on the spectrum might suggest a different range. For instance, whereas *almost no chance* refers to a probability between .01 and .05 in ICD 203, it appears to refer to a probability between 0 and .10 on the NIC spectrum. It is also noteworthy that the spectrum ranges from 0 to 100, while the probability equivalents presented in ICD 203 range from 0.01 to 0.99.

Table 18-4: ICD 203 Analytic Standard [19].

Almost No Chance	Very Unlikely	Unlikely	Roughly Even Chance	Likely	Very Likely	Almost Certain(ly)
remote	highly improbable	improbable (improbably)	roughly even odds	probable (probably)	highly probable	nearly certain
01 – 05 %	05 – 20 %	20 – 45 %	45 – 55 %	55 – 80 %	80 – 95 %	95 – 99 %

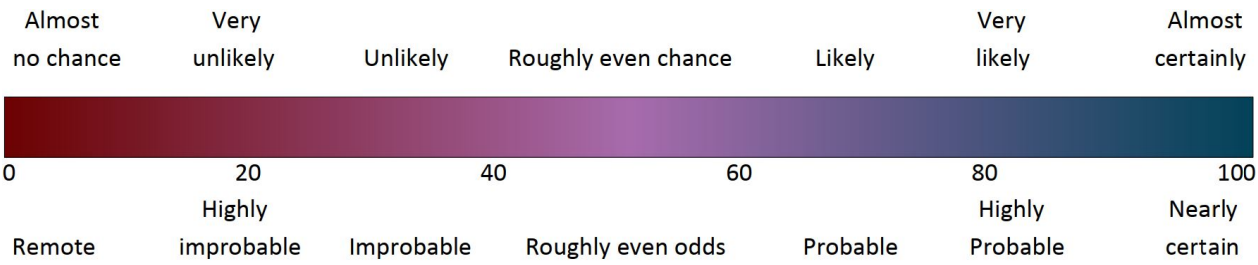


Figure 18-4: NIC Judgments of Likelihood [20].

18.2.1.4 UK Standards

The UK Defence Intelligence (DI) Uncertainty Yardstick (now superseded by the Professional Head of Intelligence Assessment [PHIA] Probability Yardstick) was mandated for use in DI assessment products, but not across the UK Intelligence Assessment Community [21]. Probability ranges were separated by gaps to prevent misrepresentation by analysts and misinterpretation by consumers. As can be seen in Table 18-5, analysts have the option of choosing from mandated synonyms. However, UK Defence Intelligence doctrine [21] provides no guidance regarding the mixing of synonyms. DI previously embedded the full conversion table in finished intelligence products for consumer reference (Table 18-5).

Table 18-5: DI Uncertainty Yardstick [21].

Qualitative Term	Associated Probability Range
Remote *or* highly unlikely	Less than 10%
Improbable *or* unlikely	15 – 20 %
Realistic probability	25 – 50 %
Probable *or* likely	55 – 70 %
Highly probable *or* highly likely	75 – 85 %
Almost certain	More than 90%

As of March 2018, use of the PHIA Probability Yardstick is mandated across the UK Intelligence Assessment Community (Table 18-6) [22]. The new standard splits *remote* and *highly unlikely* into separate levels and removes all synonyms aside from *likely*/*probably*. Verbal probability expressions are paired with percentage ranges, which can also be expressed as fractions. Percentage ranges become narrower at the extremes, and inter-range percentage gaps persist. For instance, there is no term that covers the region between a 75% and 80% chance.

Table 18-6: PHIA Probability Yardstick [22].

Probability Range	Judgement Terms	Fraction Range
≤ 5%	Remote chance	≤ 1/20
10% – 20%	Highly unlikely	1/10 – 1/5
25% – 35%	Unlikely	1/4 – 1/3
40% – < 50%	Realistic possibility	4/10 – < 1/2
55% – 75%	Likely *or* Probably	4/7 – 3/4
80% – 90%	Highly likely	4/5 – 9/10
≥ 95%	Almost certain	≥ 19/20

18.2.1.5 Norwegian Standards

Norwegian Intelligence Doctrine (*Etterretningsdoktrinen*) uses five levels to assess and communicate estimative probability (Table 18-7) [23]. These "confidence levels" (not to be confused with measures of analytic confidence) correspond to overlapping probability ranges, which are consistent with NATO doctrine. Each confidence level also contains a list of synonyms, some of which are not probabilistic (e.g., *we are convinced*; *we believe not*).

Table 18-7: Etterretningsdoktrinen Confidence Levels [23].

Confidence Levels	Synonyms
Highly likely	Highly probable. We are convinced. Virtually certain. Almost certain. High confidence. High likelihood. > 90%
Likely	Probable. We estimate. Chances are good. High-moderate confidence. 60 – 90 %
Even chance	Chances are slightly greater (or less) than even. Moderate confidence. Possible. 40 – 60 %
Unlikely	Probably not. Not likely. Improbable. We believe not. Low confidence. Possible, but not likely. 10 – 40 %
Highly unlikely	Highly improbable. Nearly impossible. Only a slight chance. Highly doubtful. < 10%

18.2.1.6 Dutch Standards

The Dutch Defence Intelligence and Security Service (DISS) Uniform Information Assessment presents verbal probability terms along a coloured spectrum (Figure 18-5) [24]. Unlike other standards examined (with the exception of NIE 2007 [18]), no probability ranges or other numerical equivalents are demarcated along the spectrum. Probability terms range from *Onwaarschijnlijk* (Improbable) to *Bevestigd* (Confirmed), which is not probabilistic. While *confirmed* could be applied to an explanatory assessment (e.g., it is confirmed that person *x* met person *y*), it is unclear how analysts could reasonably apply this term when making inherently uncertain predictions about the future (e.g., it is confirmed that person *x* will meet person *y*).

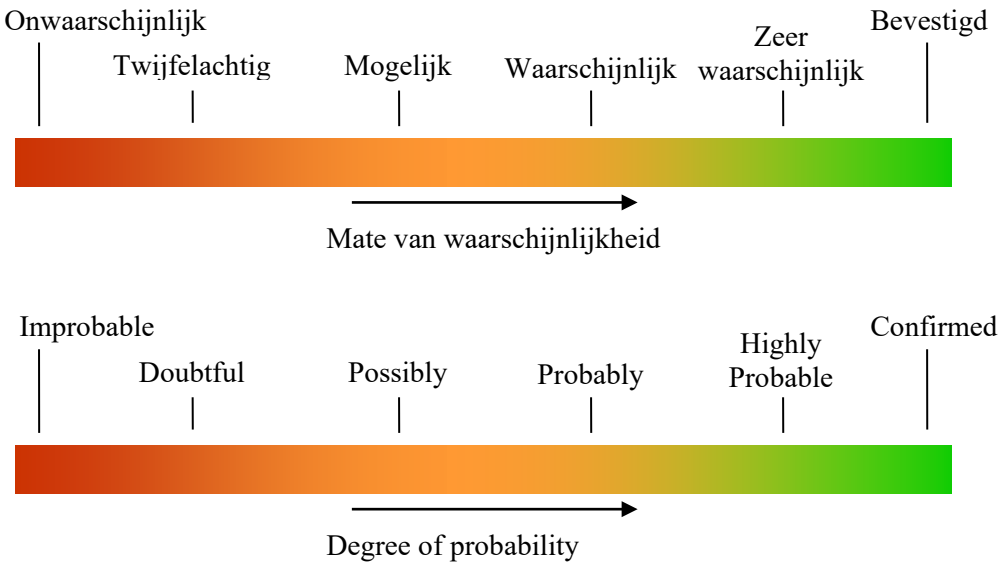


Figure 18-5: DISS Degree of Probability [24].

18.2.1.7 Danish Standards

The Danish Defence Intelligence Service (DDIS) provides the standard shown in Figure 18-6 in its 2018 Intelligence Risk Assessment [25]. The scale "does not express precise numeric differences but merely informs the reader whether something is more or less probable than something else." Brief descriptions are meant to clarify each term. However, there are cases where the description of one term contains a term from a different ordinal level (e.g., *Possible* refers to a "*likely* possibility"). See Table 18-8.

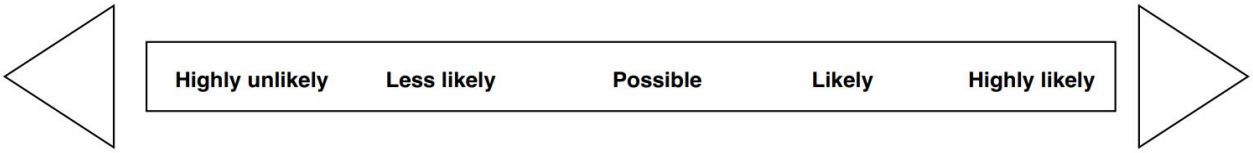


Figure 18-6: DDIS Degrees of Probability [25].

Table 18-8: DDIS Degrees of Probability [25].

Degrees of Probability
Highly unlikely: We do not expect a certain development. Such a development is (almost) not a possibility.
Less likely / doubtful: It is more likely that something will not happen than vice versa.
Possible: It is a likely possibility; however, we do not have the basis to assess whether it is more or less possible that something will happen.
Likely: It is more likely that something will happen than vice versa.
Highly likely: We expect a certain development. It has (almost) been confirmed.

18.2.2 Risk Management Standards

18.2.2.1 Canadian Risk Standards

Canadian Forces (CF) Joint Doctrine Manual: Risk Management for CF Operations measures risk likelihood on a five-level scale (Table 18-9) [26]. Probability ranges from *frequent* to *unlikely*, and is combined with the estimated consequence to determine Risk. *L*, *M*, *H*, and *E* indicate risks that are Low, Medium, High, and Extremely High, respectively. The use of *frequent*, *occasional*, and *seldom* reflect the use of this scale in gauging the frequency of recurrent phenomena, rather than the probability of one-off events [27]. The document also provides detailed descriptions of each probability phrase (Table 18-10), with different descriptions depending on the element exposed (single item, fleet or inventory of items, individual, all personnel). When assessing the risk to personnel, users estimate probability on an operation-by-operation or mission-by-mission basis. When assessing the risk to hardware, users can also estimate probability over the service life of the item(s).

Table 18-9: CF Risk Assessment Matrix [26].

Risk Assessment Matrix						
		Probability				
Severity		Frequent A	Likely B	Occasional C	Seldom D	Unlikely E
Catastrophic	I	E	E	H	H	M
Critical	II	E	H	H	M	L
Marginal	III	H	M	M	L	L
Negligible	IV	M	L	L	L	L

Table 18-10: CF Probability Categories [26].

PROBABILITY DEFINITIONS	
Element Exposed	**Definition**
FREQUENT (A): Occurs very often, continuously experienced	
Single item	Occurs very often in service life. Expected to occur several times over duration of a specific mission or operation.
Fleet or inventory of items	Occurs continuously during a specific mission or operation, or over a service life.
Individual	Occurs very often. Expected to occur several times during mission or operation.
All personnel exposed	Occurs continuously during a specific mission or operation.
LIKELY (B): Occurs several times	
Single item	Occurs several times in service life. Expected to occur during a specific mission or operation.
Fleet or inventory of items	Occurs at a high rate, but experienced intermittently (regular intervals, generally often).

PROBABILITY DEFINITIONS	
Element Exposed	**Definition**
Individual	Occurs several times. Expected to occur during a specific mission or operation.
All personnel exposed	Occurs at a high rate, but experienced intermittently.
OCCASIONAL (C): Occurs sporadically	
Single item	Occurs some time in service life. May occur about as often as not during a specific mission or operation.
Fleet or inventory of items	Occurs several times in service life.
Individual	Occurs over a period of time. May occur during a specific mission or operation, but not often.
All personnel exposed	Occurs sporadically (irregularly, sparsely, or sometimes).
SELDOM (D): Remotely possible; could occur at some time	
Single item	Occurs in service life, but only remotely possible. Not expected to occur during a specific mission or operation.
Fleet or inventory of items	Occurs as isolated incidents. Possible to occur some time in service life, but rarely. Usually does not occur.
Individual	Occurs as isolated incident. Remotely possible, but not expected to occur during a specific mission or operation.
All personnel exposed	Occurs rarely within exposed population as isolated incidents.
UNLIKELY (E): Can assume will not occur, but not impossible	
Single item	Occurrence not impossible, but can assume will almost never occur in service life. Can assume will not occur during a specific mission or operation.
Fleet or inventory of items	Occurs very rarely (almost never or improbable). Incidents may occur over service life.
Individual	Occurrence not impossible, but may assume will not occur during a specific mission or operation.
All personnel exposed	Occurs very rarely, but not impossible.

18.2.2.2 US Risk Standards

US Chairman of the Joint Chiefs of Staff Instruction (CJCSI) 3401.01E Joint Combat Capability Assessment is meant to serve as a common framework to "provide timely situational awareness of the operational and strategic risks of operations plan execution to the Joint Chiefs of Staff" [28]. The document presents a three-level likelihood scale (Table 18-11) for the Joint Staff J-2 and commander J-2 to communicate the probability of plan execution in the next 12 months. The numerical ranges are separated by gaps and vary in size. The Military Risk Matrix (Table 18-12) incorporates the likelihood of achieving strategic objectives when calculating risk. From low to high risk, the probability levels are assured, very likely, likely, and requires extraordinary measures.

Table 18-11: Joint Chiefs of Staff Intelligence/Probability Assessment [28].

Low	0 – 30 %
Medium	40 – 60 %
High	70 – 100 %

Table 18-12: Joint Staff J-5 Military Risk Assessment Matrix [28].

	Low	Moderate	Significant	High
Strategic Objective	Strategic objective: Assured	Strategic objective: Very likely	Strategic objective: Likely	Strategic objective: Requires extraordinary measures
Authorities	Full authorities provided to achieve all strategic objectives	Authorities provided to achieve most strategic objectives	Authorities are insufficient to achieve key strategic objectives	Critical authorities are not provided: ability to achieve strategic objectives is compromised
Plans	GEF/JSCP direct advanced planning: OPLANS (Level IV) or CONPLANS (Level III)	GEF/JSCP direct preliminary planning: Base Plans (Level II) or CDR's Estimate (Level I)	GEF/JSCP do not direct planning, but local plans exist or are being developed	GEF/JSCP do not direct plans and planning is not progressing
Resources	As planned	Additional resources from other plans and operations	Additional resources from other plans and operations: some significant capability shortfalls	Significant resources from other operations: some resources severely deficient or absent altogether
Resources: Timelines	As planned	Extended	Significant adjustments	Significant adjustments: may not achieve desired end-states
Resources: Unanticipated Requirements	Easily managed, minimal impact	May necessitate adjustments to plans	Will necessitate significant adjustments to plans	Unable to manage
Resources: Force Provider	Full capacity to source COCOM requirements	Can source all requirements. Worldwide force allocation solutions may result in limited duration capability gaps	Can source priority COCOM requirements. Worldwide force allocation solutions may result in extended duration capability gaps	Require full mobilization to sustain sourcing solutions to achieve strategic objectives
Resources: Services Functions, Force Management, Institutional Capacity	Full capacity to source COCOM requirements	Requires Intra-Service adjustments to source COCOM requirements	Requires joint source solutions and force substitutions to source COCOM requirements	COCOM requirements exceed Joint Force capacity to substitute capabilities

Department of Defense (DOD) Form 2977 contains the Deliberate Risk Assessment Worksheet employed by the US military [29]. DOD Form 2977 uses the same risk likelihood terms presented in CF Joint Doctrine Manual: Risk Management for CF Operations [26]. Simplified probability descriptions are embedded within the matrix, and Extremely High risk is abbreviated *EH* rather than *E* (Table 18-13). By specifying that probability is the "expected frequency" of a hazard, DOD Form 2977 directly conflates these two concepts.

Department of Defense Risk, Issue, and Opportunity Management Guide for Defense Acquisition Programs, Recommended Likelihood Criteria provides a five-level scale for assessing the likelihood of an event (Table 18-14) [30]. Qualitative terms are paired with non-overlapping probability ranges of approximately equal size. Evaluators are encouraged to estimate probability using quantitative analysis to the extent practical, but are simultaneously discouraged from expressing probabilities using fractional likelihood levels (e.g., a likelihood of 2.5) as this "incorrectly implies increased fidelity in the assessment."

Table 18-13: DOD Deliberate Risk Assessment Worksheet [28].

DELIBERATE RISK ASSESSMENT WORKSHEET						
Risk Assessment Matrix		**Probability** (*expected frequency*)				
		Frequent: Continuous, regular, or inevitable occurrences	**Likely:** Several or numerous occurrences	**Occasional:** Sporadic or intermittent occurrences	**Seldom:** Infrequent occurrences	**Unlikely:** Possible occurrences but improbable
Severity (*expected frequency*)		**A**	**B**	**C**	**D**	**E**
Catastrophic: *Death, unacceptable loss or damage, mission failure, or unit readiness eliminated*	**I**	**EH**	**EH**	**H**	**H**	**M**
Critical: *Severe injury, illness, loss, or damage; significantly degraded unit readiness or mission capability*	**II**	**EH**	**H**	**H**	**M**	**L**
Moderate: *Minor injury, illness, loss, or damage; somewhat degraded unit readiness or mission capability*	**III**	**H**	**M**	**M**	**L**	**L**
Negligible: *Minimal injury, loss, or damage; little or no impact to unit readiness or mission capability*	**IV**	**M**	**L**	**L**	**L**	**L**

Table 18-14: DOD Recommended Likelihood Criteria [30].

Level	Likelihood	Probability of Occurrence
5	Near certainty	> 80% to ≤ 99%
4	Highly likely	> 60% to ≤ 80%
3	Likely	> 40% to ≤ 60%
2	Low likelihood	> 20% to ≤ 40%
1	Not likely	> 1% to ≤ 20%

18.2.2.3 UK Risk Standards

UK Cabinet Office National Risk Register (NRR) of Civil Emergencies provides separate likelihood scales for "terrorist and other malicious attacks" (e.g., cyber attacks) and "other risks" (e.g., flooding) [31]. The probability of malicious attacks is communicated using five qualitative terms ranging from *low* to *high* (Figure 18-7), while the probability of other risks is expressed on a logarithmic scale ranging from *between 1 in 20,000 and 1 in 2,000* to *greater than 1 in 2* (Figure 18-8). Forecasts are made over a five-year timeframe, and the position of key risks within the matrices is reviewed and updated with each edition of the NRR. Depending on the scenario, analysts will use historical analysis, quantitative modelling, and scientific expertise to inform their probability estimates. The estimated likelihood of malicious attacks reflects a subjective assessment of threat actor intent and capability, plus target vulnerability. The NRR emphasizes that the two scales are not directly comparable.

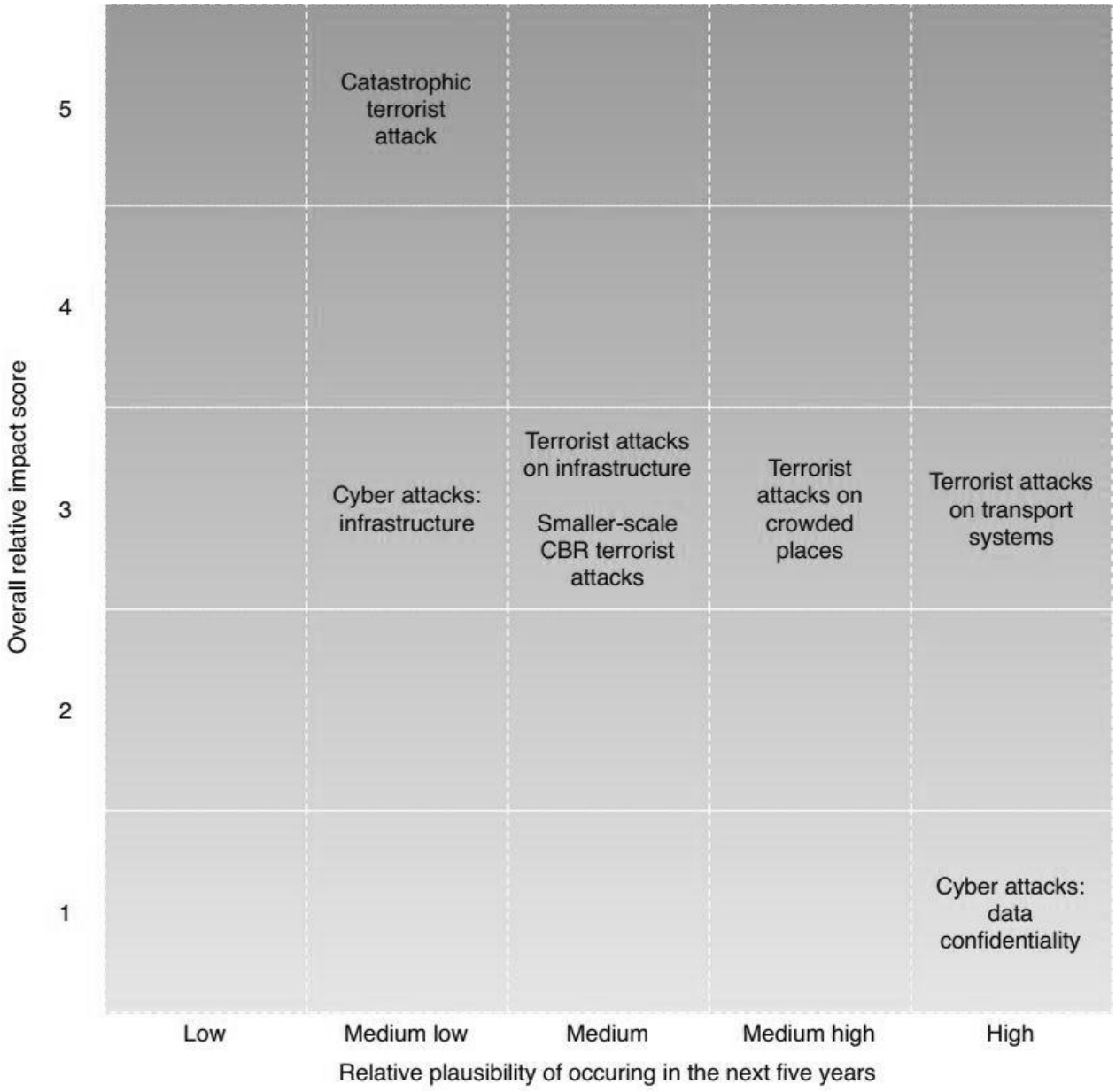


Figure 18-7: UK Cabinet Office NRR Risk of Terrorist and Other Malicious Attacks [31].

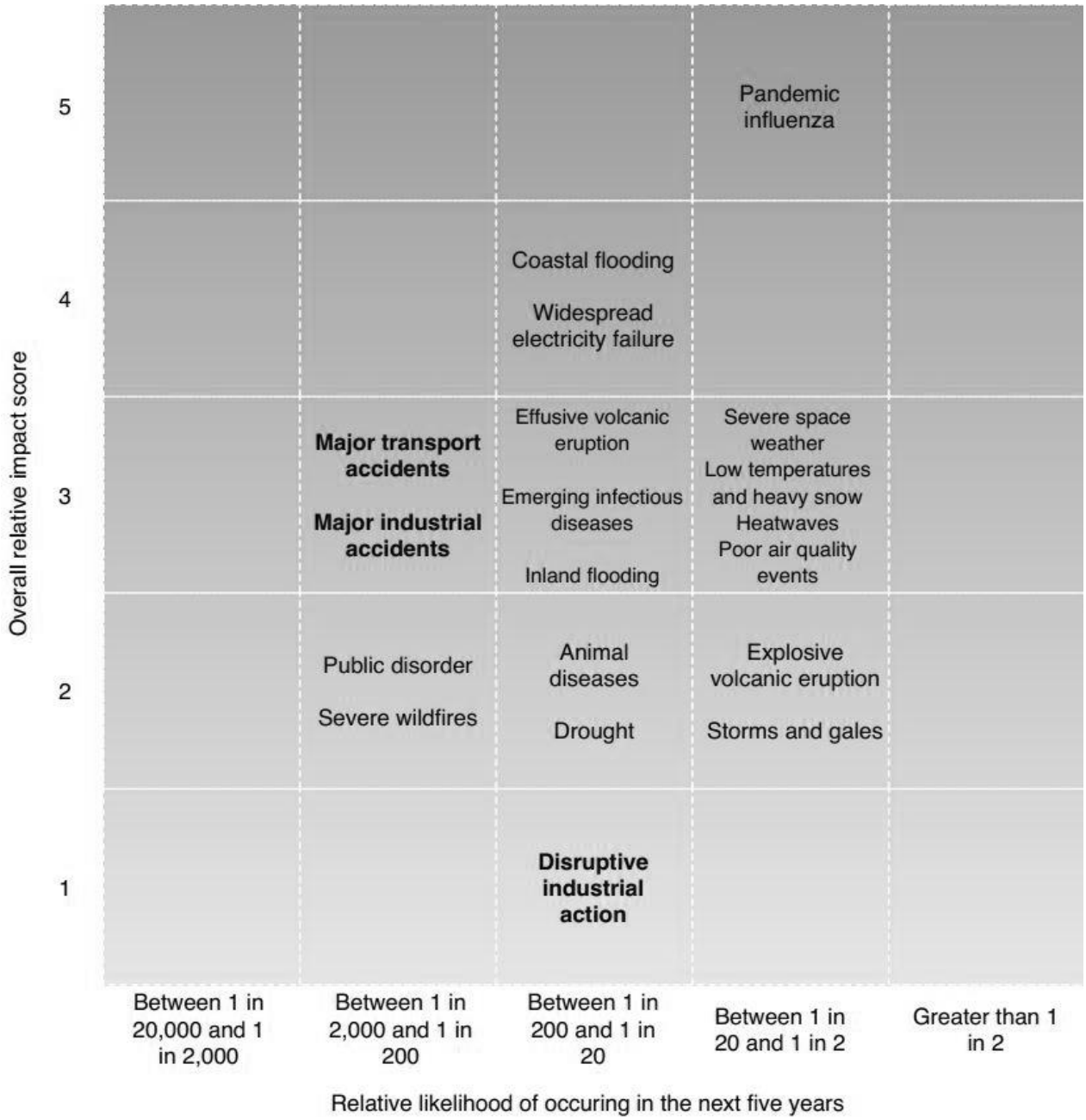


Figure 18-8: UK Cabinet Office NRR Other Risks [30].

18.2.2.4 Dutch Risk Standards

The National Safety and Security Strategy of the Netherlands is implemented across different sectors and ministries of the Dutch government to increase national resilience [32]. Separate scales are provided for evaluating the probability of "threat scenarios" (where triggers are malicious/intentional) (Table 18-19) and "hazard scenarios" (where triggers are not malicious/intentional) (Table 18-20). This dichotomy resembles the separation of "malicious attacks" and "other risks" in the UK Cabinet Office NRR [31]. In both scales, likelihood is divided into five categories (*A* to *E*), which correspond to five categories used to rate event impact (the other component of risk in this standard). For increased precision, categories *A-D* in the hazard scale are divided into three subcategories: *low*, *medium*, and *high*. Threat likelihood is expressed using

qualitative descriptions, while hazard likelihood incorporates numerical equivalents for each category and subcategory (e.g., *B – medium* = 0.1 – 0.25 %). This approach reflects that certain predictive techniques (e.g., statistical modelling) are considered more applicable to events like extreme weather (a non-malicious hazard) than terrorism (a malicious threat). While many of the scales examined become more granular at either extreme (i.e., probability ranges become more narrow with proximity to 0 or 1), the numerical values in the hazard scale are extremely precise at the low end of the scale but extremely coarse at the high end (e.g., *A* covers 0 – 0.05 %, while *E* covers 50 – 100 %). The standard justifies this lopsidedness on the grounds that it "gives a degree of robustness... that does justice to the inaccuracy of the estimation of likelihood," and because "incident scenarios will mostly cluster in the lower part of the likelihood scale." Likelihood is estimated over a five-year timeframe, as in the NRR [31].

Table 18-15: Netherlands Category Breakdown of Likelihood of Threats [32].

Category	Qualitative Threat Description
A	No concrete indication and the event is not deemed conceivable.
B	No concrete indication; event is deemed far-fetched but conceivable.
C	No concrete indication, but event is conceivable.
D	The event is deemed very conceivable.
E	Concrete indication that the event will take place.

Table 18-16: Netherlands Category Breakdown of Likelihood of Hazards [32].

Category	% per 5 Years		Quantitative (%)	Qualitative Description of the Hazard
A	< 0.05	A – low A – medium A – high	< 0.005 0.005 – 0.02 0.02 – 0.05	Very unlikely
B	0.05 – 0.5	B – low B – medium B – high	0.05 – 0.1 0.1 – 0.25 0.25 – 0.5	Unlikely
C	0.5 – 5	C – low C – medium C – high	0.5 – 1 1 – 2.5 2.5 – 5	Possible
D	5 – 50	D – low D – medium D – high	5 – 10 10 – 25 25 – 50	Likely
E	50 – 100	E	50 – 100	Very likely

18.2.3 Standards Used in Other Domains

US Global Change Research Program Climate Science Special Report (USGCRP CSSR) Fifth-Order Draft provides estimative probability standards for communicating forecasts related to global climate change (Table 18-17) [33]. The standard was adapted from previous USGCRP assessments [34], [35], and the Intergovernmental Panel on Climate Change's (IPCC's) Fifth Assessment Report [36]. Where the probability ranges described previously tend to overlap only at rating boundaries (e.g., 45 – 55 %, 55 – 80 %), overlap in the USGCRP CSSR standard is much more substantial. For instance, the equivalent probability range of *Likely* (66% – 100%) subsumes the ranges of *Very Likely* (90% – 100%), *Extremely Likely* (95% – 100%), and *Virtually Certain* (99% – 100%), while overlapping at the boundary with *About as Likely as Not* (33% – 66%). Comparable overlap can be observed in most of the IPCC standards outlined below.

Table 18-17: USGCRP CSSR 5OD Likelihood [33].

Likelihood
Virtually certain 99% – 100%
Extremely likely 95% – 100%
Very likely 90% – 100%
Likely 66% – 100%
About as likely as not 33% – 66%
Unlikely 0% – 33%
Very unlikely 0% – 10%
Extremely unlikely 0% – 5%
Exceptionally unlikely 0% – 1%

The IPCC Third Assessment Report (TAR) contains two different estimative probability scales used by Working Group I (WGI) (Table 18-18) and Working Group II (WGII) (Table 18-19), respectively [37]. The scales pair qualitative terms with probability ranges of varying size. Both WGI and WGII use overlapping ranges in their general estimates. The scales used in IPCC TAR range between five and seven probability levels, with some variation in terminology.

Table 18-18: IPCC TAR WGI Likelihood Statements [37].

Terminology	Likelihood of the Occurrence/Outcome
Virtually certain	Greater than 99% chance a result is true
Very likely	90 – 99 % chance
Likely	66 – 90 % chance
Medium likelihood	33 – 66 % chance
Unlikely	10 – 33 % chance
Very unlikely	1 – 10 % chance
Exceptionally unlikely	Less than 1% chance

Table 18-19: IPCC TAR WGII Confidence Statements [37].

Terminology	Likelihood of the Occurrence/Outcome
Very high	95% or greater
High	67 – 95 %
Medium	33 – 67 %
Low	5 – 33 %
Very low	5% or less

The IPCC Fourth [37] and Fifth Assessment Reports [35] (AR4 and AR5, respectively) provide estimative probability standards with between five and ten levels. All of the probability ranges overlap and vary in size. See Table 18-20, Table 18-21, Table 18-22 and Table 18-23.

Table 18-20: IPCC AR4 (WGI)/AR5 (WGII) Likelihood Terminology [36], [38].

Likelihood Terminology	Likelihood of the Occurrence/Outcome
Virtually certain	> 99% probability
Extremely likely	> 95% probability
Very likely	> 90% probability
Likely	> 66% probability
More likely than not	> 50% probability
About as likely as not	33 to 66 % probability
Unlikely	< 33% probability
Very unlikely	< 10% probability
Extremely unlikely	< 5% probability
Exceptionally unlikely	< 1% probability

Table 18-21: IPCC AR4 (WGI)/AR5 (WGII) Description of Likelihood [38].

Terminology	Likelihood of the Occurrence/Outcome
Virtually certain	> 99% probability of occurrence
Very likely	90 to 99 % probability
Likely	66 to 90 % probability
About as likely as not	33 to 66 % probability
Unlikely	10 to 33 % probability
Very unlikely	1 to 10 % probability
Exceptionally unlikely	< 1% probability

Table 18-22: IPCC AR4 (WGI) Special Report on Emissions Scenarios Likelihood [38].

Terminology	Likelihood of the Occurrence/Outcome
Virtually certain	> 99% probability of occurrence
Very extremely likely	> 95%
Very likely	> 90%
Likely	> 66%
More likely than not	> 50%

Table 18-23: IPCC AR5 (WGI, WGII) Likelihood Scale [36].

Term	Likelihood of Outcome
Virtually certain	99 – 100 % probability
Very likely	90 – 100 % probability
Likely	66 – 100 % probability
About as likely as not	33 – 66 % probability
Unlikely	0 – 33 % probability
Very unlikely	0 – 10 % probability
Exceptionally unlikely	0 – 1 % probability

The PBL Netherlands Environmental Assessment Agency Guide for Uncertainty Communication [39] / Dutch Environmental Balance [40] reproduces the terminology and probability ranges employed by IPCC AR4 Working Group II [38], with the addition of colour coding (Table 18-24). Terminology is provided in both English and Dutch.

Table 18-24: Netherlands Environmental Assessment Agency Verbal Information [39], [40].

Dutch Term	English Synonym (IPCC)	Likelihood	Colour Code Tables
Nagenoeg zeker	Virtually certain	> 99%	
Zeer waarschijnlijk	Very likely	90 – 99 %	
Waarschijnlijk	Likely	66 – 99 %	
Fifty-fifty; circa 50%	About as likely as not (new) Medium likelihood (old)	33 – 66 %	
Onwaarschijnlijk	Unlikely	10 – 33 %	
Zeer onwaarschijnlijk	Very unlikely	1 – 10 %	
Nagenoeg uitgesloten	Exceptionally unlikely	< 1%	

The European Food Safety Authority's Scientific Committee Guidance on Uncertainty in EFSA Scientific Assessment provides a seven-level estimative probability standard to aid in expert knowledge elicitation (Table 18-25) [41]. The scale is based on IPCC standards, but has slightly different terminology (e.g., *extremely unlikely* is used in the place of *exceptionally unlikely*). The document encourages experts to communicate judgements as a precise probability or range of probabilities without using the EFSA standard when possible.

Table 18-25: EFSA Probability Guidance [41].

Probability Term	Subjective Probability Range
Extremely likely	99 – 100 %
Very likely	90 – 99 %
Likely	66 – 90 %
As likely as not	33 – 66 %
Unlikely	10 – 33 %
Very unlikely	1 – 10 %
Extremely unlikely	0 – 1 %

18.3 TERMINOLOGICAL ISSUES

Among the standards examined, estimative probability is generally communicated using verbal probability terms (e.g., *likely*, *improbable*). Under the former IAS MEA standard, analysts were required to assign numerical probability values on a 9-point scale, but these were omitted from finished intelligence products [2]. As the title of our chapter indicates, the use of verbal probability terms is problematic first and foremost because they are inherently vague. That is, interpretations of verbal probabilities are shown to vary widely between individuals (i.e., different individuals assign different numerical equivalents to the same phrase), with considerable overlap among the numerical ranges assigned to certain terms [3], [42], [43], [44], [45], [46], [47], [48], [49], [50]. Studies also reveal significant within-subject variability; that is, a single individual will assign different numerical equivalents to the same phrase [45], [51].

Interpretations can vary depending on context [3], [4], [52], [53], [54], [55] and presentation [49], [56], [57], as well as individual interpreter characteristics such as numeracy, language, and personal attitudes [58], [59], [60]. Of particular relevance to multinational organizations like NATO, native English speakers are shown to interpret and assign numerical values to verbal probability terms differently than individuals who speak a first language other than English [60]. Studies also show that verbal-to-numeric translations vary significantly between cultural groups, as does probabilistic thinking more broadly [61], [62], [63], [64].

In terms of specific verbiage, many of the standards examined are structured around the use of mirror-image terms (e.g., *likely*/*unlikely*, *highly likely* / *highly unlikely*). However, studies have repeatedly shown that numerical interpretations of mirror-image terms are not symmetrical (i.e., the means and medians of positive values tend to be closer to the midpoint; *likely* is closer to 50% than *unlikely*), nor do they usually sum to 100% [42], [46], [47], [49]. All else being equal, negatively worded probability terms are interpreted with greater variability than positively worded terms [65].

Several standards also incorporate synonyms, ostensibly to facilitate stylistic flexibility (e.g., *likely* or *probably*; *very unlikely* or *highly improbable*) [2], [16], [17], [19], [20], [21]. However, as noted previously, ICD 203 is the only document examined with explicit instructions for mixing and matching these terms [19]. Specifically, analysts are discouraged from mixing terms and must include a disclaimer in products where mixing occurs. Without clear guidance, consumers may interpret synonyms inconsistently. For instance, the 2004 Review of Intelligence on Weapons of Mass Destruction prepared for the UK House of Commons notes that British policymakers assumed different synonyms had distinct meanings, while analysts claimed they were simply employing natural language [66], [67].

With the exception of the former IAS MEA standard, which carefully mapped synonyms to numerical probabilities based on the linguistic probability literature [2], synonyms provided in the standards examined have not been evaluated for equivalence, which may increase the likelihood of miscommunication. Whether two or more terms are synonymous is an empirically verifiable question. Mandel [4] confirmed that most of the terms used as synonyms in the former IAS MEA standard were close to one another in average interpretation. For instance, whereas the median numerical probability that best represents *likely* was 75%, the median probability of *probable* was 71% (see also Ref. [68]). The same method could be used to assess synonymic equivalence in other scales, and to evaluate whether interpretations of these terms vary significantly between analysts and consumers (especially decision makers with limited exposure to intelligence products).

It is also noteworthy that some terms are treated synonymously in one standard but treated as different degrees of probability in another standard. These inconsistencies could impede interoperability and precipitate miscommunication between analysts familiar with different standards. For instance, *remote chance* and *highly unlikely* are treated as different degrees of probability in ICD 203 [19], but as synonyms in the former DI Uncertainty Yardstick [21]. The recently implemented PHIA Probability Yardstick [22], which supersedes the DI Uncertainty Yardstick, separates *remote* and *highly unlikely*, perhaps in the interest of fostering consistency with the US standard. However, this may represent a step backwards, as Ho *et al.* [68] found that these terms had near-perfectly synonymous interpretations by intelligence analysts. It is possible that when the stipulated ranges are provided the meaning of the terms in either case is better anchored.

The inclusion of midpoint terms such as *fifty-fifty*, *even chance*, and *circa 50%*, may be problematic, given that individuals often misapply these expressions to communicate their subjective uncertainty or an inability to assess, rather than a 50% probability of the target event occurring (or not occurring) [69]. If analysts default to these fence-sitting "agnostic" phrases, they are essentially representing a given situation as a coin toss, and therefore doing little to support decision making. For this reason, Kesselman [70] advocates the use of *chances a little better (or less)* to ensure that estimates reduce uncertainty (if only marginally), although this strikes us as a poor solution. Event probabilities should be assigned without *ad hoc* fixes likes these, and

analysts should be encouraged to have a better grasp of different types of uncertainty that impact their judgements. DIA Tradecraft Note 01-15 [17] recommends that analysts revaluate key assumptions, and conduct additional collection and analysis, if they arrive in the middle probability "gray zone." However, this advice also fails to differentiate cases in which the gray zone is the result of poor evidence or knowledge or where it represents a reasonable summary of the evidence, which may entail conflicting sources that on balance warrant a middle-of-the-road assessment. In the end, these instructions and injunctions may do more to foster unreliability within and across analysts than to improve the process of assessing and communicating uncertainty.

Analysts should be instructed to accurately characterize uncertainty and to use the midpoint of the probability scale for the right reasons; namely, when that is what the evidence points to as the best probabilistic assessment [71]. If an analyst is completely unsure due to a paucity of relevant or credible information, and hence faces utter epistemic uncertainty, then .5 is not appropriate. The analyst should refrain from giving an estimate and explain that any estimate would have very low confidence attached to it. On the other hand, if credible evidence exists, and when weighed in its totality it points to a probability of .5, then that should unreservedly be the assessment given (e.g., if an analyst concludes that abundant, high-quality evidence indicates an extremely close electoral race). Under such circumstances, we do not think analysts should be prompted to tip the scales slightly one way or the other. More generally, we do not advocate for any method that routinely or even periodically encourages analysts to express anything other than what they regard as their best probability estimate for the event being judged.

Another terminological issue observed in several standards is the inclusion of so-called weasel words, such as *might*, *may*, and *could* (e.g., see Refs. [16], [17], [24]). These vague and highly ambiguous phrases are particularly easy for consumers to misinterpret [6]. Furthermore, individuals are shown to exploit ambiguous probability phrases to reach self-serving conclusions [72]. The use of weasel words may thus encourage consumers to interpret estimates in ways that are politically expedient.

Weasel words may also reduce analyst accountability by impeding post-mortem analysis [67], [73]. For instance, an estimate that an event *could* happen cannot be found definitively 'wrong,' thereby limiting the extent to which analysts and researchers can evaluate its accuracy and provide feedback. Many of the forecasts that Mandel and Barnes [13], [14] excluded from their quantitative analyses of accuracy were removed for this reason. Javorsek and Schwitz [74] argue that "only by explicitly identifying the probabilities of possible outcomes associated with a measure of significant consequence can we evaluate and improve our intelligence process while signalling the probability of catastrophic events."

Compounding the use of vague terminology is the inclusion of expressions that are not probabilistic. As noted previously, the DISS Uniform Information Assessment [24] uses *confirmed* as its highest probability level, while the Norwegian Intelligence Doctrine [23] incorporates phrases such as *we are convinced* and *we believe not*. In certain risk assessment standards, probability expressions are mixed with frequency terms. As described by Mandel [27], under the CF Joint Doctrine Manual: Risk Management for CF Operations [26], likelihood is rated *frequent* (a frequency), *likely* (a probability), *occasional* (a frequency), *seldom* (a frequency), or *unlikely* (a probability). Similar issues are present in DOD Form 2977 [29].

Whereas the terminology described above is overly ambiguous, the inclusion of terms indicating certainty, such as is *certain*, *certainly*, *will*, *will not*, and *no prospect*, may also be problematic, given the estimative and inherently uncertain nature of intelligence. Generally, an event that has no chance of occurring will not be analyzed, and an event deemed certain will be analyzed in terms of its potential ramifications, but not its probability [67]. For these reasons, NATO AJP-2.1 [12] explicitly discourages expressions of certainty. Similarly, DIA Tradecraft Note 01-15 [17] recommends that analysts using terms of certainty review whether their judgements are simply statements of obvious facts. Once again, we do not advocate that intelligence organizations establish such *ad hoc* rules for assessing probabilities. Analysts should be encouraged to report as accurately as they can. If that means estimating a probability of 0, .5, or 1, then that

is what they should report. However, this speaks to the inadequacy of language for conveying extreme probabilities. Bayesian reasoners tend to steer clear or 0 or 1 because these values imply that no amount of new evidence can change one's mind, but they may be inclined to say that the likelihood is one in a million or 999,999 in a million, which still leaves a chance for belief change. There are, however, no probability terms that allow us to effectively distinguish one in a million from even one in a hundred, and maybe even one in ten. Vague verbiage simply does not reinforce rigorous thinking since not only are analysts not encouraged to practice precision; they are actually denied the opportunity.

A final terminological consideration worth discussing is the extent to which standards incorporate the verbal uncertainty expressions that analysts and consumers actually use to communicate estimative probability. Dhami [10] asked 26 UK intelligence analysts to provide probability terms they would use to represent a variety of 10% intervals. While the analysts provided a total of 160 unique probability phrases, *realistic possibility* (the midpoint expression used in the new PHIA Probability Yardstick [22]), *roughly even chance*, and *roughly even odds* (the two midpoint expressions used in ICD 203 [19]) were among phrases which did not appear in any of the analysts' lexicons. This feature of certain standards may inadvertently encourage analysts to avoid large portions of uncertainty scales, or to use unsanctioned probability terms instead [10].

18.4 SCORING ISSUES

Many of the standards examined link verbal probability phrases to prescribed numerical equivalents (e.g., see Refs. [2], [12], [19]). However, a conversion guide is not always provided to consumers, meaning their interpretations may deviate significantly from the analyst's intent. Even when standards mandate the inclusion of these guides in finished products (e.g., Refs. [19], [22]), it is unclear to what extent consumers reference them as they interpret analytic findings. In a study on uncertainty standards employed by the IPCC, Budescu *et al.* [75] find that access to a conversion guide only slightly improves the consistency of interpretations, and that embedding numerical values directly into estimates is more effective.

With the exception of the former IAS MEA standard [2], none of the standards examined effectively leverage research on verbal probability interpretations. As noted above, Ho *et al.* [68] find that certain synonyms incorporated into the NIC [20] and UK JDP [21] standards are numerically indistinguishable (i.e., there is genuine synonymic equivalence), however, user interpretations of the phrases do not align with the numerical ranges stipulated [10]. Similarly, Budescu *et al.* [58], [59], [75] find that public interpretations of probability terms used by the IPCC are highly regressive (i.e., high probabilities are underestimated, while low probabilities are overestimated) compared to the numerical equivalents outlined in the conversion guide. As demonstrated by Ho *et al.* [68], standards could benefit from correcting numerical equivalents to accurately reflect how verbal probability terms are interpreted by analysts and consumers.

NIE 2007 [18], DISS Uniform Information Assessment [24], and DIA Tradecraft Note 01-15 [17] forgo the use of numerical values altogether, as do most of the risk assessment standards examined. Standards may omit numerical values due to the assumption that quantitative expressions lead consumers to be overconfident and overly risk seeking in their subsequent decision making. However, contrary to the first assumption, Friedman, Lerner, and Zeckhauser [9] find that national security officials presented numerical probability assessments were in fact less confident in supporting proposed actions and they were more amenable to additional information gathering.

As noted previously, numerical probabilities may also enable analysts to better characterize the likelihood of low probability, high consequence events. Terms like *remote chance* simply cannot be used to differentiate spanning orders of magnitude such as one in a million, one in a thousand, and one percent [71]. Moreover, while pairing verbal probability terms with numerical ranges in text can be an effective means of reducing variation in consumer interpretation (e.g., Ref. [75]), such pairings have been studied with stipulated ranges rather than those estimated by the individuals making the judgements, and stipulated ranges are typically

unhelpful for communicating low probability events since they span orders upon orders of magnitude. For instance, if *remote chance* has a stipulated range of "less than 10%", it cannot convey to decision makers whether they are in a "one in a million" situation or a "one percent" situation. Such precision may not be important in many decision-making scenarios, but it surely is important in at least some highly consequential ones (namely, reasoning about low probability, high consequence severity outcomes).

Further undermining communication fidelity, several standards incorporate overlapping ranges, which may encourage analysts to obscure probability, or facilitate biased interpretation by consumers. For instance, in NATO AJP-2.1 [12], *likely* refers to a 60 – 90 % probability, while *even chance* refers to 40 – 60 %. Depending on how a forecast is presented, consumers may interpret a 60% probability as *likely* rather than *even chance*, or vice versa. UK JDP 2-00 [21] employed non-overlapping, non-continuous ranges (e.g., 15 – 20 %, 25 – 50 %, 55 – 70 %) to prevent intentional or unintentional blurring of probability assessments. However, this feature is also problematic as it may force analysts to distort their estimates to comply with the scale [20].

As the foregoing observations hopefully make clear, there is no ideal solution to the challenge of communicating uncertainty in intelligence that relies on the use of verbal probabilities either in conjunction with numerical ranges or on their own. Even if verbal probability standards were grounded in empirical evidence, as in the case of Barnes [2], studies indicate that most people struggle to suppress the meanings they normally associate with such terms (e.g., see Refs. [10], [52]). In spite of this, for over half a century, intelligence communities continue to balk at the suggestion of using numerical probabilities to convey event likelihood. Instead, they devise and tinker with standards for using vague verbiage. They do so in ways that are just different enough over time and across organizations so that inconsistency in usage is almost guaranteed. Realizing the limitations of verbal probabilities, intelligence organizations further institute *ad hoc* injunctions to fix various problems, which in our view create others that are possibly worse. The net result is that the IC has created something like a Rube Goldberg machine for expressing probabilities in their intelligence products. However, at least Rube Goldberg machines do a fair job. We believe that, in principle, the same cannot be said for the IC's current communication standards.

Finally, we note that we are not entirely unsympathetic to the IC's resistance to using numerical probabilities. Clearly, in many contexts it feels unnatural, if not unnecessary, to forgo the use of words to communicate likelihood. Such vague usage may often merely signal a rough indication of intent, as in "I'm fairly likely to go to the party on Saturday." Precision requires additional mental energy. It is harder to decide that one is 79% likely to go to a party than to judge it "fairly likely". Not only is the latter imprecise, it is imprecise about its imprecision, which is to say that a speaker uttering this phrase will be unlikely to have a confidence interval in mind. To the speaker, "fairly likely" may mean little more than "I think I will go, but I am not sure." If this level of precision suffices, then we cannot object to the use of vague verbiage, and there are theoretical arguments in favor of such allowances (see Ref. [76]). We have little doubt that some situations, perhaps even in the realm of intelligence, are of this type. However, we do not believe that the vast majority of cases in intelligence production are of this type. Nor does it seem plausible to us that the cases where greater precision and clarity are warranted are so few in number and collectively so inconsequential as to merit the continuation of the current vague verbiage approach. The IC may want to factor in pathways for quickly communicating at coarse levels where only the most rudimentary indications are needed, but such pathways should be no more than off ramps of an otherwise high-fidelity communication freeway built on a quantitative infrastructure for the communication of primarily subjective probabilities.

18.5 REFERENCES

[1] Irwin, D., and Mandel, D.R. (2018). Methods for Communicating Estimative Probability in Intelligence to Decision Makers: An Annotated Collection. DRDC Scientific Letter DRDC-RDDC-2018-L017.

[2] Barnes, A. (2016). Making intelligence analysis more intelligent: Using numeric probabilities. *Intelligence and National Security* 31, (3):327-344.

[3] Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting,* 1(3):257-269.

[4] Mandel, D.R. (2015). Accuracy of intelligence forecasts from the intelligence consumer's perspective. *Policy Insights from the Behavioral and Brain Sciences,* 2(1):111-120.

[5] Friedman, J.A., Baker, J.D., Mellers, B.A., Tetlock, P.E., and Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly*, 62(2):410-422.

[6] Kent, S. (1964). Words of estimative probability. *Studies in Intelligence,* 8(4):49-65.

[7] Wheaton, K.J. (2012). The revolution begins on page five: The changing nature of NIEs. *International Journal of Intelligence and CounterIntelligence,* 25(2):330-349.

[8] Marchio, J. (2014). Analytic tradecraft and the intelligence community: Enduring value, intermittent emphasis. *Intelligence and National Security,* 29(2):159-183.

[9] Friedman, J.A., Lerner, J.S., and Zeckhauser, R. (2017). Behavioral consequences of probabilistic precision: Experimental evidence from national security professionals. *International Organization*, 71(4):803-826.

[10] Dhami, M.K. (2018). Towards an evidence-based approach to communicating uncertainty in intelligence analysis. *Intelligence and National Security,* 33(2):257-272.

[11] Dhami, M.K., Mandel, D.R., Mellers, B.A., and Tetlock, P.E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science,* 10(6):753-757.

[12] North Atlantic Treaty Organization. (2016). *Allied Joint Doctrine for Intelligence Procedures AJP-2.1*. Brussels, Belgium.

[13] Mandel, D.R., and Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences of the United States of America,* 111(30):10984-10989.

[14] Mandel, D.R., and Barnes, A. (2018). Geopolitical forecasting skill in strategic intelligence: Geopolitical forecasting skill. *Journal of Behavioral Decision Making,* 31(1):127-137.

[15] Mandel, D.R. (2019). Can decision science improve intelligence analysis? In: *Researching National Security Intelligence: Multidisciplinary Approaches,* Coulthart, S., Landon-Murray, M., and Van Puyvelde, D. (Eds.). 117-140. Washington DC: Georgetown University Press.

[16] Canadian Forces Intelligence Command. (2015). *Aide-Mémoire on Intelligence Analysis Tradecraft* (v 6.0). Ottawa, ON: DND.

[17] Defense Intelligence Agency. (2015). *Expressing Analytic Certainty*. Tradecraft Note 01-15. Washington DC: Defense Intelligence Agency.

[18] Office of the Director of National Intelligence. (2007). *Iran: Nuclear Intentions and Capabilities*. National Intelligence Estimate. Washington DC. Retrieved from https://www.dni.gov/files/documents/Newsroom/Reports%20and%20Pubs/20071203_release.pdf.

[19] Office of the Director of National Intelligence. (2015). *Intelligence Community Directive ICD 203: Analytic Standards*. Retrieved from https://fas.org/irp/dni/icd/icd-203.pdf.

[20] Office of the Director of National Intelligence. (2017). *Intelligence Community Assessment – Assessing Russian Activities and Intentions in Recent US Elections*. Retrieved from https://www.dni.gov/files/documents/ICA_2017_01.pdf.

[21] United Kingdom Ministry of Defence. (2011). *Joint Doctrine Publication JDP 2-00, Understanding and Intelligence Support to Joint Operations*, 3rd ed. Swindon, UK. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/311572/20110830_jdp2_00_ed3_with_change1.pdf.

[22] UK Defence Intelligence. (n.d.). *PHIA Probability Yardstick.* London, UK.

[23] Norwegian Ministry of Defence. (2013). *Etterretningsdoktrinen*. Oslo, Norway.

[24] Anonymous interviewee. (2015). *Dutch Ministry of Defence Uniform Information Assessment*.

[25] Danish Defence Intelligence Service. (2018). *Intelligence Risk Assessment 2018: An Assessment of Developments Abroad Impacting Danish Security*. Copenhagen, Denmark. Retrieved from https://fe-ddis.dk/SiteCollectionDocuments/FE/EfterretningsmaessigeRisikovurderinger/Risk_Assessment2018.pdf.

[26] Department of National Defence. (2007). *Risk Management for CF Operations*. Ottawa, ON: DND.

[27] Mandel, D.R. (2007). *Toward a Concept of Risk for Effective Military Decision Making*. DRDC Toronto Technical Report 2007-124. Toronto, ON: DRDC.

[28] United States Joint Chiefs of Staff. (2010). *Chairman of the Joint Chiefs of Staff Instruction CJCSI 3401.01E Joint Combat Capability Assessment*. Washington DC: United States Joint Chiefs of Staff.

[29] Department of Defense. (2017). *Form 2977: Risk Assessment Worksheet*. Washington DC. Retrieved from http://www.acq.osd.mil/se/docs/2017-RIO.pdf.

[30] Office of the Deputy Assistant Secretary of Defense for Systems Engineering. (2015). *Department of Defense Risk, Issue, and Opportunity Management Guide for Defense Acquisitions Programs*. Washington DC. Retrieved from http://bbp.dau.mil/docs/RIO-Guide-Jun2015.pdf.

[31] United Kingdom Cabinet Office. (2015). *National Risk Register of Civil Emergencies*. London, UK. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/419549/20150331_2015-NRR-WA_Final.pdf.

[32] Netherlands Ministry of Security and Justice. (2009). *Working with Scenarios, Risk Assessment and Capabilities in the National Safety and Security Strategy of the Netherlands*. The Hague, the Netherlands. Retrieved from https://english.nctv.nl/binaries/working-with-scenarios-risk-assessment-and-capabilities_tcm32-84297.pdf.

[33] Wuebbles, D.J., Fahey, D.W., Hibbard, K.A., Dokken, D.J., Stewart, B.C., and Maycock, T.K. (Eds.). (2017). *Climate Science Special Report: A Sustained Assessment Activity of the U.S. Global Change Research Program*. Washington DC: US Global Change Research Program. Retrieved from https://unfccc.int/sites/default/files/resource/84_CSSR2017_Executive_Summary_Q1-3.pdf.

[34] Melillo, J.M., Richmond, T.C., and Yohe, G.W. (Eds.). (2014). *Climate Change Impacts in the United States: The Third National Climate Assessment*. Washington DC: US Global Change Research Program.

[35] United States Global Change Research Program. (2016). *The Impacts of Climate Change on Human Health in the United States: A Scientific Assessment*. Retrieved from https://health2016.globalchange.gov/.

[36] Intergovernmental Panel on Climate Change. (2014). *Fifth Assessment Report AR5 2014*. Cambridge, UK and New York City, NY. Retrieved from: https://www.ipcc.ch/report/ar5/

[37] Intergovernmental Panel on Climate Change. (2001). *Third Assessment Report TAR 2001*. Cambridge, UK and New York City, NY. Retrieved from https://www.ipcc.ch/ipccreports/tar/.

[38] Intergovernmental Panel on Climate Change. (2007). *Fourth Assessment Report AR4 2007* Cambridge, UK and New York City, NY. Retrieved from https://www.ipcc.ch/report/ar4/.

[39] PBL Netherlands Environmental Assessment Agency. (2013). *Guide for Uncertainty Communication*. The Hague, the Netherlands. Retrieved from http://www.pbl.nl/sites/default/files/cms/publicaties/PBL_2013_Guide-for-uncertainty-communication_1339.pdf.

[40] PBL Netherlands Environmental Assessment Agency. (2006). *Environmental Balance 2006*. The Hague, the Netherlands. Retrieved from http://www.pbl.nl/sites/default/files/cms/publicaties/opm_summary_mb2006.pdf.

[41] European Food Safety Authority Scientific Committee. (n.d.). *Guidance on Uncertainty in EFSA Scientific Assessment*. Revised draft for internal testing. Parma, Italy. Retrieved from https://www.efsa.europa.eu/sites/default/files/consultation/150618.pdf.

[42] Lichtenstein, S., and Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science,* 9(10):563-564.

[43] Hakel, M.D. (1968). How often is often? *American Psychologist,* 23(7):533-534.

[44] Johnson, E.M. (1973). *Numerical Encoding of Qualitative Expressions of Uncertainty.* Technical Paper 250. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

[45] Budescu, D.V., and Wallsten, T.S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes,* 36 (3):391-405.

[46] Chesley, G.R. (1985). Interpretation of uncertainty expressions. *Contemporary Accounting Research,* 2(2):179-199.

[47] Reagan, R.T., Mosteller, F., and Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of Applied Psychology,* 74(3):433-442.

[48] Mullet, E., and Rivet, I. (1991). Comprehension of verbal probability expressions in children and adolescents. *Language and Communication,* 11(3):217-225.

[49] Clarke, V.A., Ruffin, C.L., Hill, D.J., and Beamen, A.L. (1992). Ratings of orally presented verbal expressions of probability by a heterogeneous sample. *Journal of Applied Social Psychology*, 22(8):638-656.

[50] Teigen, K.H., and Brun, W. (1995). Yes, but it is uncertain: Direction and communicative intention of verbal probabilistic terms. *Acta Psychologica,* 88(3):233-258.

[51] Budescu, D.V., and Wallsten, T.S. (1990). Dyadic decisions with numerical and verbal probabilities. *Organizational Behavior and Human Decision Processes,* 46(2):240-263.

[52] Wallsten, T.S., Budescu, D.V., Rapoport, A., Zwick, R., and Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General,* 115(4):348-365.

[53] Brun, W., and Teigen, K.H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes,* 41(3):390-404.

[54] Witteman, C., and Renooij, S. (2003). Evaluation of a verbal-numerical probability scale. *International Journal of Approximate Reasoning,* 33(2):117-131.

[55] Patt, A.G., and Schrag, D.P. (2003). Using specific language to describe risk and probability. *Climatic Change,* 61(1):17-30.

[56] Hamm, R.M. (1991). Selection of verbal probabilities: A solution for some problems of verbal probability expression. *Organizational Behavior and Human Decision Processes,* 48(2):193-223.

[57] Bergenstrom, A., and Sherr, L. (2003). The effect of order of presentation of verbal probability expressions on numerical estimates in a medical context. *Psychology, Health & Medicine*, 8(4):391-398.

[58] Budescu, D.V., Por, H., and Broomell, S.B. (2012). Effective communication of uncertainty in the IPCC reports. *Climatic Change,* 113(2):181-200.

[59] Budescu, D.V., Por, H., Broomell, S.B., and Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change,* 4(6):508-512.

[60] Brooks, Z.S. (2016). Bilingual decision making: Verbal probability, ethics, and cognition. Doctoral dissertation. Tucson, AZ: The University of Arizona.

[61] Doupnik, T.S., and Richter, M. (2004). The impact of culture on the interpretation of “in context” verbal probability expressions. *Journal of International Accounting Research,* 3(1):1-20.

[62] Lau, L., and Ranyard, R. (2005). Chinese and English probabilistic thinking and risk taking in gambling. *Journal of Cross-Cultural Psychology,* 36(5):621-627.

[63] Doupnik, T.S., and Riccio, E.L. (2006). The influence of conservatism and secrecy on the interpretation of verbal probability expressions in the anglo and latin cultural areas. *The International Journal of Accounting,* 41(3):237.

[64] Li-Jun, J., and Megan, K. (2013). Judgement and decision making across cultures: Judgement and decision making across cultures. *Advances in Psychological Science,* 21(3):381-388.

[65] Smithson, M., Budescu, D.V., Broomell, S.B., and Por, H. (2012). Never say "not": Impact of negative wording in probability phrases on imprecise probability judgments. *International Journal of Approximate Reasoning,* 53(8):1262.

[66] Butler, F.E.R., Chilcot, J., Inge, P.A., Mates, M., and Taylor, A. (2004). *Review of Intelligence on Weapons of Mass Destruction.* The Butler Review, HC 898. London, UK. Retrieved from http://news.bbc.co.uk/nol/shared/bsp/hi/pdfs/14_07_04_butler.pdf.

[67] Lowenthal, M.M. (2012). *Intelligence: From Secrets to Policy*, 5th ed. Washington DC: CQ Press.

[68] Ho, E., Budescu, D.V., Dhami, M.K., and Mandel, D.R. (2015). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science & Policy,* 1(2):43-55.

[69] Fischhoff, B., and Bruine de Bruin, W.J.A.W. (1999). Fifty-fifty = 50%? *Journal of Behavioral Decision Making,* 12(2):149-163.

[70] Kesselman, R.F. (2008). Verbal probability expressions in national intelligence estimates: A comprehensive analysis of trends from the fifties to post 9/11. Master's thesis. Erie, PA: Mercyhurst College.

[71] Friedman, J.A., and Zeckhauser, R. (2012). Assessing uncertainty in intelligence. *Intelligence and National Security*, 27(6):824-847.

[72] Piercey, M.D. (2009). Motivated reasoning and verbal vs. numerical probability assessment: Evidence from an accounting context. *Organizational Behavior and Human Decision Processes*, 108(2):330-341.

[73] Snider, L.B. (2008). *The Agency and the Hill: CIA's Relationship with Congress*, 1946-2004. Washington DC: Central Intelligence Agency, Center for the Study of Intelligence.

[74] Javorsek, D., II, and Schwitz, J.G. (2014). Probing uncertainty, complexity, and human agency in intelligence. *Intelligence and National Security,* 29(5):639-653.

[75] Budescu, D.V., Broomell, S., and Por, H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science,* 20(3):299-308.

[76] Wallsten, T.S., and Budescu, D.V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *The Knowledge Engineering Review,* 10(1):43-62.



Chapter 19 – HOW INTELLIGENCE ORGANIZATIONS COMMUNICATE CONFIDENCE (UNCLEARLY)[1, 2]

Daniel Irwin and David R. Mandel
Defence Research and Development Canada
CANADA

19.1 INTRODUCTION

Given that intelligence is typically derived from incomplete and ambiguous evidence, analysts must accurately assess and communicate their level of uncertainty to consumers [2]. One facet of this perennial challenge is the communication of analytic confidence, or the level of confidence that an analyst has in his or her judgements, including those already qualified by probability terms such as "very unlikely" or "almost certainly". Consumers are better equipped to make sound decisions when they understand the methodological and evidential strength (or flimsiness) of intelligence assessments. Effective communication of confidence also militates against the pernicious misconception that the Intelligence Community (IC) is omniscient [2].

As part of broader efforts to improve communication fidelity and rein in subjectivity in intelligence production, most intelligence organizations have adopted standardized lexicons for rating and communicating analytic confidence. These standards provide a range of confidence levels (e.g., *high*, *moderate*, *low*), along with relevant rating criteria, and are often paired with scales used to express estimative probability (for a review of estimative probability standards, see chapters by Irwin and Mandel and by Dhami and Mandel in this report).

Notwithstanding these standardization efforts, there is evidence that expressions of confidence are easily misinterpreted by consumers. For instance, during a 2013 US congressional hearing, Rep. Doug Lamborn inadvertently disclosed a line from a Intelligence Agency (DIA) assessment stating that "DIA assesses with moderate confidence that North [Korea] currently has nuclear weapons capable of delivery by ballistic missiles" [3]. According to DIA standards from the time, *moderate* confidence reflected "partially corroborated information from good sources, several assumptions, and/or a mixture of strong and weak inferences" [4]. Despite this guidance, several participants in the hearing dismissed the assessment outright, believing that any judgements with less than *high* confidence could be ignored [5].

There is also evidence that the terms stipulated in confidence standards are misunderstood (or at least misapplied) by intelligence practitioners. In the 2007 National Intelligence Estimate (NIE 2007) Iran: Nuclear Intentions and Capabilities, probabilistic expressions (e.g., *probably*, *unlikely*) are said to "reflect the [Intelligence] Community's estimates of the likelihood of developments or events" [6]. Meanwhile, confidence levels indicate the extent to which "assessments and estimates are supported by information that varies in scope, quality and sourcing." Despite being prominently featured in the product, this guidance does not correspond to how the concepts of likelihood and confidence are actually applied. In the declassified Key Judgments section, there are 19 instances where 'confidence' is used to express likelihood, rather than qualify it (e.g., "We judge with high confidence that in fall 2003, Tehran halted its nuclear weapons program." [3]) [7]. At no point in the Key Judgments section are expressions of probability and confidence used to jointly characterize uncertainty as officially intended [7]. This apparent misapplication of analytic confidence is particularly troubling given that the NIE is a high-level strategic intelligence product, representing the coordinated judgement of the entire US IC.

[1] This chapter expands on the following client-focused document: Irwin and Mandel [1].

[2] Funding support for this work is provided by the Canadian Safety and Security Program Project CSSP-2016-TI-2224 (Improving Intelligence Assessment Processes with Decision Science).

[3] Note: A similar misapplication of confidence occurs in the DIA estimate described previously.

In this chapter, we review a non-exhaustive collection of the analytic confidence standards compiled by members and affiliates of NATO's SAS-114 Research Task Group on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making. These include standards used in intelligence production and other domains (e.g., climate science). We outline common problematic features that might compromise efforts to support decision making, while identifying avenues for future research and development. Ultimately, we argue that current confidence standards are poorly conceived, ambiguous, vague, and unclear, and may effectively augment the potential for miscommunication, which the IC seeks to mitigate. We recommend ways to improve confidence scales in their current form, but we also propose a more dramatic overhaul involving the use of numerical probabilities.

19.2 OVERVIEW OF CURRENT STANDARDS

19.2.1 National Security Intelligence Standards

19.2.1.1 NATO Standards

NATO Allied Joint Doctrine for Intelligence Procedures (NATO AJP-2.1) outlines confidence communication procedures intended for use by NATO members as well as external partners [8]. Analytic confidence is assessed on a qualitative, three-level scale, and is associated with information credibility, source reliability, correlation, and number of collection capabilities utilized. NATO doctrine emphasizes that "throughout interpretation and all-source fusion, the analyst should attempt to find confirming information or intelligence," as opposed to seeking out disconfirming information, or attempting to strike a balance between confirming and disconfirming information. It is also noteworthy that, where other confidence standards incorporate analytic considerations like subject matter expertise and methodological rigour (e.g., Refs. [9], [10], [11]), NATO doctrine focuses exclusively on evidentiary characteristics (see Table 19-1).

Table 19-1: NATO AJP-2.1 2016 Confidence Levels [8].

Confidence Levels	
High	Good quality of information, evidence from multiple collection capabilities, possible to make a clear judgement.
Moderate	Evidence is open to a number of interpretations, or is credible and plausible but lacks correlation.
Low	Fragmentary information, or from collection capabilities of dubious reliability.

19.2.1.2 Canadian Standards

Canadian Forces Intelligence Command (CFINTCOM) Aide Memoire on Intelligence Analysis Tradecraft [9] presents a three-level confidence scale adapted from DIA Tradecraft Note 01-15 [10]. In line with the DIA Tradecraft Note, likelihood and confidence are conceptualized as separate components of "Analytic Certainty". "Where appropriate," CFINTCOM instructs analysts to clearly indicate both their level of confidence and their reasons for ascribing it. Analytic confidence is based on three main factors:

- **Evidence:** "the strength of the knowledge base, to include the quality of the evidence and our depth of understanding about the issue."
- **Assumptions:** "the number and importance of assumptions used to fill information gaps."
- **Reasoning:** "the strength of the logic underpinning the argument, which encompasses the number and strength of analytic inferences as well as the rigour of the analytic methodology applied to the product."

Analysts are discouraged from using *Complete Confidence* and *No Confidence*; complete confidence likely means stating the obvious, while no confidence means the analyst will probably not be commenting on the subject. Analysts are instructed to communicate confidence to consumers as explicitly as possible, but are simultaneously told that they "will not [need to] very often." Despite distinguishing between analytic confidence and estimative probability, the document lists *possible*, *may*, *might*, *could*, *can*, *perhaps*, *unsure*, *unknown*, *do not know*, *unable to* assess, and *undetermined* as estimative phrases to use when analytic confidence is low. Analysts are expected to outline their confidence ratings in a dedicated textbox, or to integrate them into the narrative text of the product (see Table 19-2).

Table 19-2: CFINTCOM Confidence [9].

Confidence	
High	Well-corroborated information from proven sources, low potential for deception, non-critical assumptions and/or gaps, or undisputed reasoning.
Moderate	Partially corroborated information from good sources, moderate potential for deception, potentially critical assumptions used to fill gaps, or a mix of inferences.
Low	Uncorroborated information from good or marginal sources, high potential for deception, key assumptions used to fill critical gaps, or mostly weak inferences.

The Public Safety Canada All Hazards Risk Assessment Methodology Guideline 2012 – 2013 [12] was developed by Public Safety Canada and the Defence Research and Development Canada Centre for Security Science as a guideline for federal government institutions in Canada. Analysts are expected to provide analytic confidence for both Likelihood Analysis and Impact/Consequence Analysis. The scale uses five levels, rather than three. Analytic confidence is based on knowledge of the issue, information credibility, information volume, and consistency. While the number/importance of assumptions is not explicitly factored into the scale, analysts are instructed to clearly outline any assumptions made. When presented to decision makers, expressions of *low* or *very low* confidence should be expected to "necessitate caution in interpretation," while expressions of *high* or *very high* confidence should be expected to "inspire trust in immediate actions required to treat pressing risks" (see Table 19-3).

Table 19-3: Public Safety Canada Confidence Level [12].

Confidence Level	Quantification
A	Very high confidence in the judgement based on a thorough knowledge of the issue, the very large quantity and quality of relevant data and totally consistent relevant assessments.
B	High confidence in the judgement based on a very large body of knowledge on the issue, the large quantity and quality of relevant data and very consistent relevant assessments.
C	Moderate confidence in the judgement based on a considerable body of knowledge on the issue, the considerable quantity and quality of relevant data and consistent relevant assessments.
D	Low confidence in the judgement based on a relatively small body of knowledge on the issue, the relatively small quantity and quality of relevant data and somewhat consistent relevant assessments.
E	Very Low confidence in the judgement based on small to insignificant body of knowledge on the issue, quantity and quality of relevant data and/or inconsistent relevant assessments.

19.2.1.3 US Standards

NIE 2007 describes analytic confidence in its "What We Mean When We Say" explanation of estimative language [6]. Analytic confidence is connected to information credibility, plausibility, corroboration, and the "nature of the issue." The document emphasizes that *high confidence* judgements still carry the risk of being wrong. As discussed in the authors' chapter on estimative probability, NIE 2007 directly links estimative probability and analytic confidence by prescribing the use of certain probability terms (e.g., *might*, *may*) under conditions of *low confidence*, where "relevant information is unavailable, sketchy, or fragmented" (see Table 19-4).

Table 19-4: NIE 2007 Confidence in Assessments [5].

High Confidence generally indicates that our judgments are based on high-quality information, and/or that the nature of the issue makes it possible to render a solid judgment. A "high confidence" judgment is not a fact or a certainty, however, and such judgments still carry a risk of being wrong.
Moderate Confidence generally means that the information is credibly sourced and plausible but not of sufficient quality or not corroborated sufficiently to warrant a higher level of confidence.
Low Confidence generally means that the information's credibility and/or plausibility is questionable, or that the information is too fragmented or poorly corroborated to make solid analytic inferences, or that we have significant concerns or problems with the sources.

US Joint Chiefs of Staff Joint Publication (JP) 2-0 Joint Intelligence [11] is meant to guide joint and multinational intelligence activities across the spectrum of US military operations. Unlike other standards examined, JP 2-0 nests words of estimative probability directly under analytic confidence ratings, combining two concepts that are typically separated. For instance, under conditions of *Moderate* confidence, analysts are told to communicate their estimates using the phrases *likely*, *unlikely*, *probable*, *improbable*, *anticipate*, or *appear*. Analytic confidence rests on information credibility, source reliability, corroboration, methodological rigour, logic, intelligence gaps, and number of assumptions. Analysts are instructed to assess each factor independently, and then in concert with other factors. Different judgements within a product may have varying levels of confidence. See Table 19-5.

Table 19-5: JP 2-0 Confidence in Analytic Judgments [11].

Expressing Confidence in Analytic Judgment		
Low	**Moderate**	**High**
• Uncorroborated information from good or marginal sources • Many assumptions • Mostly weak logical inferences, minimal methods application • Glaring intelligence gaps exist	• Partially corroborated information from good sources • Several assumptions • Mix of strong and weak inferences and methods • Minimum intelligence gaps exist	• Well-corroborated information from proven sources • Minimal assumptions • Strong logical inferences and methods • No or minor intelligence gaps exist
Terms/Expressions	**Terms/Expressions**	**Terms/Expressions**
• Possible • Could, may, might • Cannot judge, unclear	• Likely, unlikely • Probably, improbable • Anticipate, appear	• Will, will not • Almost certainly, remote • Highly likely, highly unlikely • Expect, assert, affirm

DIA Tradecraft Note 01-15 [10] references a 13 June 2013 ODNI memo stating that the estimated probability of an event must be distinguished from analytic confidence. As noted above, DIA conceptualizes likelihood and confidence as ostensibly separate components of "Analytic Certainty." Confidence is based on information credibility, information volume, source reliability, source diversity, subject matter expertise, potential for deception, importance of assumptions, use of Structured Analytic Techniques (SATs), and the persuasiveness of the analyst's arguments. Analysts are required to indicate their analytic confidence in the "Source Summary and Confidence Level Statement" embedded in finished intelligence products, but can also make in-text references to confidence. While the document stresses the separation of estimative probability and analytic confidence, it proscribes the use of unlikely, likely, improbable, and probably to communicate likelihood when analytic confidence is low. As with the CFINTCOM standard [9], analysts are discouraged from using Complete Confidence and No Confidence, despite these options being incorporated into the scale (Figure 19-1). The "Expressing Analytic Certainty" graphic (Figure 19-2) is embedded in DIA products for consumer reference.

Confidence level factors include: Sourcing, Potential for Deception, Assumptions, Gaps, and Reasoning (consider all elements within the identified level before making a final determination. Do not assign a confidence level for exploratory or alternative analysis	
Complete Confidence	**Totally reliable and corroborated information with no assumptions and clear, undisputed reasoning; or contextual/historical, or other knowledge base foundation to support this level.** *Complete confidence determinations show not generally be used to communicate analytic judgment.*
High	• Well-corroborated information from proven sources, extensive databases, and/or historical understanding of the issue, or subject matter expertise • Low potential for deception exists • Assumptions used to fill gaps are not critical to the analysis • Reasoning dominated by strong logical inferences developed through multiple analytic techniques or an established, repeatable methodology
Moderate	• Partially corroborated information from good sources (a mix of proven and semi-proven sources who have been fairly accurate) with some databases and/or historical understanding of the issue, or subject matter expertise • Moderate potential for deception exists • Assumptions potentially critical to the analysis are used to fill gaps • Reasoning with a mixture of strong and weak inferences developed through simple analytic techniques or an established methodology
Low	• Uncorroborated information from good or marginal sources (a mix of semi-proven sources that have been somewhat accurate and/or new, unproven sources) with minimal databases; historical understanding of the issue, or subject matter expertise • High potential for deception exists • Key assumptions critical to the analysis are used to fill gaps • Reasoning dominated by weak inferences developed through few analytic techniques or methodologies
No Confidence	**No direct information to support the assessment. Products with no confidence should be confined to exploratory analysis with a scope note and strong methodology statement to support the purpose and logic used to construct the product.** **Do not express "no confidence" levels in the text. Logic, reasoning, and references to similar historical events provide the basis for the scenario depicted. Indicators forecasting the scenario are mandatory.** *"No Confidence" determinations should not generally be used to communicate analytic judgments.*

Figure 19-1: DIA Identifying Confidence [10].

National Intelligence Council (NIC) Confidence in the Sources Supporting Judgements appears in Annex B of Intelligence Community Assessment 2017-01D [13], and illustrates how confidence levels are explained to intelligence consumers under ICD 203. Confidence levels are said to "provide assessments of the quality and quantity of the source information that supports judgements." In other words, the NIC confidence standard is exclusively focused on evidentiary characteristics. The NIC standard closely resembles NIE 2007 Confidence in Assessments, but omits "the nature of the issue" as a confidence determinant (perhaps to maximize the focus on information quality) (see Table 19-6).

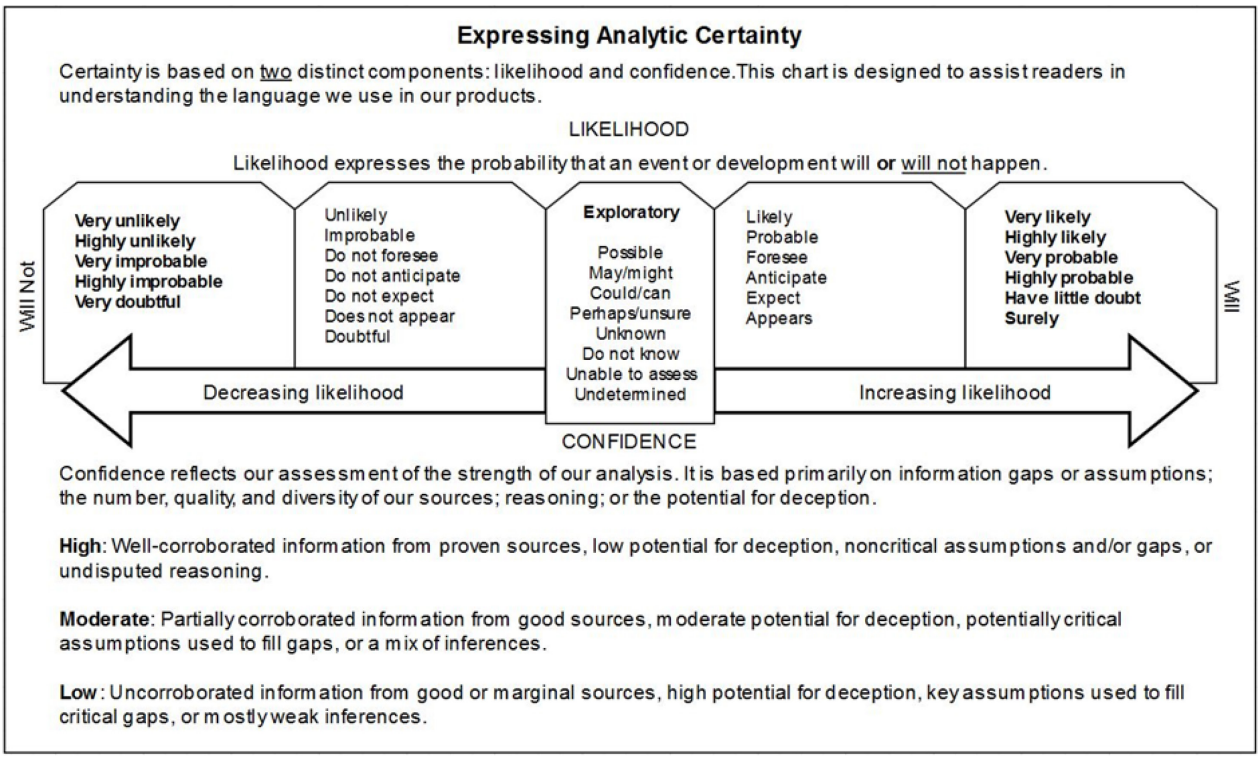


Figure 19-2: DIA Expressing Analytic Certainty [10].

Table 19-6: NIC Confidence in the Sources Supporting Judgments [13].

High confidence generally indicates that judgments are based on high-quality information from multiple sources. High confidence in a judgment does not imply that the assessment is a fact or a certainty; such judgments might be wrong.
Moderate confidence generally means that the information is credibly sourced and plausible but not of sufficient quality or not corroborated sufficiently to warrant a higher level of confidence.
Low confidence generally means that the information's credibility and/or plausibility is uncertain, that the information is too fragmented or poorly corroborated to make solid analytic inferences, or that reliability of the sources is questionable.

19.2.2 Standards Used in Other Domains

US Global Change Research Program Climate Science Special Report (USGCRP CSSR) Fourth National Climate Assessment (Volume 1) [14] provides a standard for communicating analytic confidence related to global climate change forecasts (Table 19-7). The standard was adapted from previous USGCRP assessments and the Intergovernmental Panel on Climate Change's (IPCC) Fifth Assessment Report [15]. Analytic confidence is rated on a four-level scale, based on evidentiary characteristics (e.g., source diversity, consistency) as well as the degree of consensus between experts and any models utilized. Similar attention to evidence quality and expert/model consensus can be observed in most of the IPCC standards outlined below. Unlike some of the IPCC standards examined (e.g., Refs. [15], [16]), the USGCRP CSSR standard forgoes the use of numerical values.

Table 19-7: USGCRP CSSR Confidence Level [14].

Confidence Level
Very High
Strong evidence (established theory, multiple sources, consistent results, well-documented and accepted methods, etc.), high consensus
High
Moderate evidence (several sources, some consistency, methods vary and/or documentation limited, etc.), medium consensus
Medium
Suggestive evidence (a few sources, limited consistency, models incomplete, methods emerging, etc.), competing schools of thought
Low
Inconclusive evidence (limited sources, extrapolations, inconsistent findings, poor documentation and/or methods not tested, etc.), disagreement or lack of opinions among experts

Guidance Papers on the Cross-Cutting Issues of the Third Assessment Report (TAR) of the Intergovernmental Panel on Climate Change [17] proposes guidelines for communicating analytic confidence in the IPCC TAR. The proposed "Scale for Assessing State of Knowledge" includes numerical values with overlapping ranges of varying widths (see Table 19-8). The document also proposes a 2 by 2 matrix of confidence terms and provides an example of radar plots / snowflake charts used to communicate confidence.

Table 19-8: IPCC Scale for Assessing State of Knowledge [17].

(1.00) "**Very High Confidence**" (0.95)
(0.95) "**High Confidence**" (0.67)
(0.67) "**Medium Confidence**" (0.33)
(0.33) "**Low Confidence**" (0.05)
(0.05) "**Very Low Confidence**" (0.00)

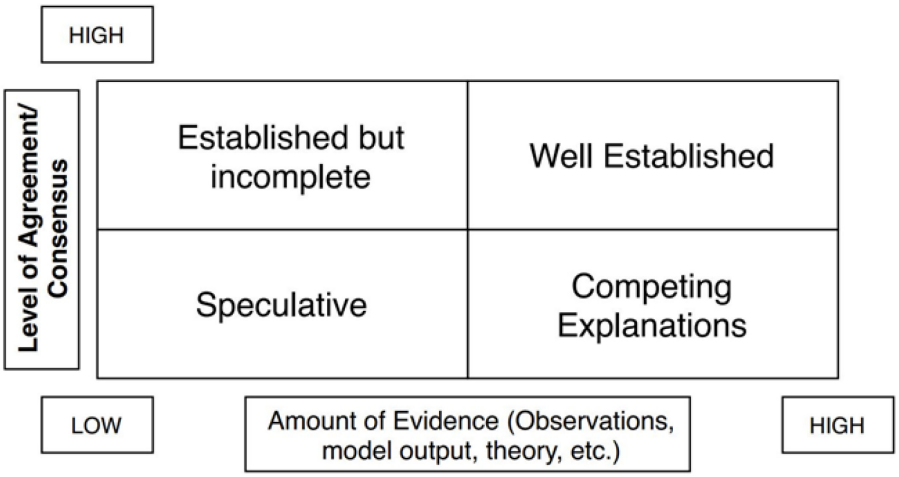


Figure 19-3: IPCC Supplemental Qualitative Uncertainty Terms [17].

The IPCC TAR Working Group I (WGI) confidence standard [18] is based on the three-level system proposed by Moss and Schneider [19]. Confidence reflects the degree of consensus amongst modellers, as well as the quantity of evidence available to support the finding (see Table 19-9).

Table 19-9: IPCC TAR WGI Levels of Confidence [18].

Well-established	Nearly all models behave the same way; observations are consistent with nearly all models; systematic experiments conducted with many models support the finding.
Evolving	Some models support the finding; different models account for different aspects of the observations; different aspects of key processes can be invoked to support the finding.
Speculative	Conceptually plausible idea that has only been tried in one model or has very large uncertainties associated with it.

The IPCC TAR Working Group II (WGII) "Qualitative Assessment of the State of Knowledge" resembles the 2 by 2 matrix proposed in the IPCC TAR Guidance Papers [20]. "Level of scientific understanding" is similar to analytic confidence and paired with estimative probability. Level of scientific understanding reflects the volume of evidence and the level of agreement/consistency across analytic models and observations. See Table 19-10.

Intergovernmental Panel on Climate Change Guidance Notes for Lead Authors of the IPCC Fourth Assessment Report (AR4) on Addressing Uncertainties [21] proposes an analytic confidence scale with non-overlapping numerical values (Table 19-11). A proposed 3 by 3 confidence matrix expands on the matrix proposed in the IPCC TAR Guidance Papers [17], but omits qualitative labels for medium evidence/agreement values. See Figure 19-4.

IPCC Fourth and Fifth Assessment Reports [15], [16] highlight evidence volume, quality, and agreement (across observations and models) as the key determinants of analytic confidence. Quantitative measures of confidence are used for analytic conclusions derived using quantitative methodologies. See Table 19-12, Table 19-13, Figure 19-5 and Figure 19-6.

World Anti-Doping Agency (WADA) Information Gathering and Intelligence Sharing Guidelines [22] uses six estimative probability terms and numerical values to express the likelihood that an intelligence assessment is true (i.e., analytic confidence). Aside from the highest confidence level ("Confirmed"), the percentage ranges are the same size and do not overlap. See Table 19-14.

Table 19-10: IPCC TAR WGII Qualitative Assessment of the State of Knowledge [20].

Well-Established	Models incorporate known processes, observations are consistent with models, or multiple lines of evidence support the finding.
Established but Incomplete	Models incorporate most known processes, although some parameterizations may not be well tested; observations are somewhat consistent but incomplete; current empirical estimates are well founded, but the possibility of changes in governing processes over time is considerable; or only one or a few lines of evidence support the finding.
Competing Explanations	Different model representations account for different aspects of observations or evidence or incorporate different aspects of key processes, leading to competing explanations.
Speculative	Conceptually plausible ideas that are not adequately represented in the literature or that contain many difficult-to-reduce uncertainties.

Table 19-11: IPCC AR4 Guidance Notes Quantitatively Calibrated Levels of Confidence [21].

Terminology	Degree of Confidence in Being Correct
Very high confidence	At least 9 out of 10 chance of being correct
High confidence	About 8 out of 10 chance
Medium confidence	About 5 out of 10 chance
Low confidence	About 2 out of 10 chance
Very low confidence	Less than 1 out of 10 chance

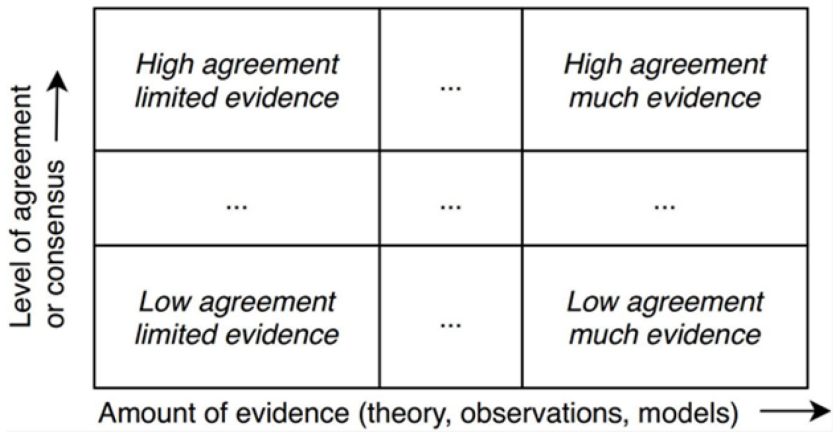


Figure 19-4: IPCC AR4 Guidance Notes Quantitatively Defined Levels of Understanding [21].

Table 19-12: IPCC AR4/AR5 Description of Confidence (WGI and WGII) [15], [16].

Terminology	Degree of Confidence in Being Correct
Very high confidence	At least 9 out of 10 chance of being correct
High confidence	About 8 out of 10 chance
Medium confidence	About 5 out of 10 chance
Low confidence	About 2 out of 10 chance
Very low confidence	Less than 1 out of 10 chance

Table 19-13: IPCC AR5 Confidence (WGI and WGII) [15].

Very high confidence
High confidence
Medium confidence
Low confidence
Very low confidence

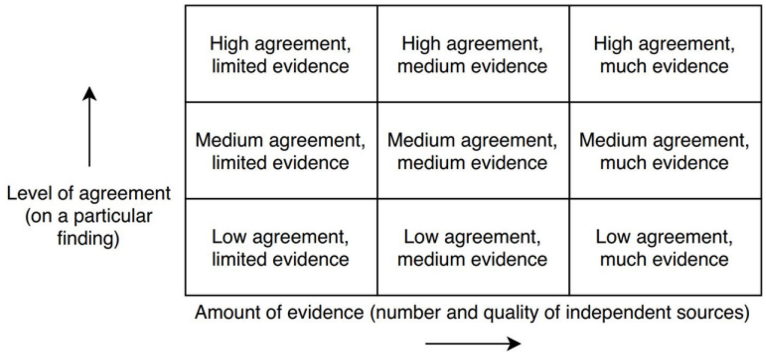


Figure 19-5: IPCC AR4 WGIII Qualitative Definition of Uncertainty [16].

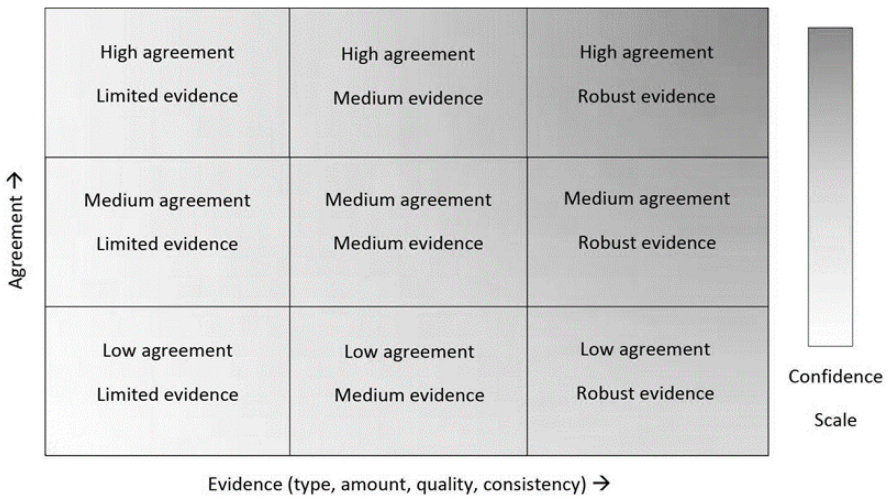


Figure 19-6: IPCC AR5 WGI Confidence [15].

Table 19-14: WADA Levels of Confidence [22].

Assessing Confidence	
Term Used	**Level of Confidence**
Confirmed	100%
Probable	80% – 99%
Likely	60% – 79%
Possible	40% – 59%
Unlikely	20% – 39%
Very unlikely	Less than 20%

19.3 TERMINOLOGICAL ISSUES

The analytic confidence standards examined generally incorporate the following determinants:

- Source reliability;
- Information credibility;
- Evidence consistency/convergence;
- Strength of logic/reasoning; and
- Quantity and significance of assumptions and information gaps.

Few standards attempt to operationalize these determinants or outline formal mechanisms for evaluation. Instead, they tend to provide vague, qualitative descriptions for each confidence level, which may lead to inconsistent confidence assessments. For example, under US JP 2-0 [11], analysts are expected to characterize the logic of their own inferences as "weak," "a mix of strong and weak," or "strong," without specific evaluation criteria. "A mix of strong and weak" is particularly ambiguous, given that the phrase encapsulates cases where inferences are more "weak" than "strong" and vice versa. Yet, a recipient of this phrase might infer some distribution over these two types. Perhaps the mention of both suggests to some individuals that the two types are about equally balanced.

Another example of vague terminology occurs in Public Safety Canada's All Hazards Risk Assessment Methodology [12], which bases confidence, in part, on whether relevant assessments are "totally consistent," "very consistent," "consistent," "somewhat consistent," or "inconsistent." In this case, inclusion of the unmodified term ("consistent") is problematic because it subsumes three of the four modified terms. Moreover, the meaning of terms such as "very", "totally", and "somewhat" are fuzzy and will vary across individuals as well as within individuals across different pragmatic contexts.

Along with the strength of inferences, US JP 2-0 [11] partially bases analytic confidence on whether "glaring intelligence gaps exist," "minimum intelligence gaps exist," or "no or minor intelligence gaps exist." Interpretations of "glaring," "minimum," and "minor" will likely vary between analysts, particularly because these terms do not form an intuitive progression; analysts may presume the difference between "glaring" and "minimum" is considerable, while the difference between "minimum" and "minor" appears narrow, if not highly ambiguous. The phrase "no or minor intelligence gaps" is also problematic because intelligence professionals rarely, if ever, produce estimates with no intelligence gaps. Use of this phrase may thus contribute to overconfidence or false expectations among consumers.

Inconsistencies stemming from this vague terminology are likely compounded by the lack of mechanisms for self-assessment. Analysts are unlikely to fully recognize information gaps and incorporate them into confidence evaluations, particularly when responding to highly complex requirements [23]. CFINTCOM [9] provides instructions for a Key Assumptions Check, which involves identifying and scoring each assumption based on its relevance and support (i.e., identifying information gaps). However, this scoring scheme ultimately relies on subjective judgements, and lacks an empirical basis. Furthermore, the Key Assumptions Check is listed with other SATs, and not directly incorporated into the analytic confidence scale or its instructional text. Future research could explore methods to reliably identify assumptions and information gaps, along with other relevant confidence determinants. At the very least, experiments could evaluate how analysts and consumers interpret current qualitative descriptors (e.g., how they subjectively quantify "glaring" versus "minor" intelligence gaps) to generate sets of less ambiguous terminology.

19.4 CONVERGENCE

Issues may also arise from the emphasis most standards place on evidence convergence as a determinant of analytic confidence. Convergence can help eliminate false assumptions and false/deceptive information, but may not necessarily prevent analysts from deriving high confidence from outdated information [24]. Under current standards, a large body of highly credible and consistent information could contribute to high analytic confidence, despite being out of date. A possible solution would be to incorporate a measure of information "recency" [24]. In assessing recency, the relevant time period will vary depending on the scope of the intelligence requirement. For instance, information several months old will not necessarily detract from confidence in a strategic-level forecast of a target country's military development. In contrast, information several days old may lower confidence in a tactical-level assessment of enemy troop movements. Future research could develop context-appropriate methods of evaluating recency and integrate them into confidence standards. With consideration of recency, care should be taken to ensure that analysts do not derive excessive confidence from information simply by virtue of it being current.

The emphasis on convergence may also lead analysts to inflate their confidence by accumulating seemingly useful but redundant information. None of the standards examined provide guidance regarding what level of convergence warrants each confidence rating. After an early point, information has a negligible impact on predictive accuracy, but can inadvertently boost confidence when analysts consider redundant pieces of information to be individually valid [23], [25], [26], [27]. A related issue is what Heuer [23] describes as the "law of small numbers," or the tendency of analysts to overlook a small sample when the available evidence is consistent. If analysts cannot determine that the evidence is representative of the total body of potentially available information, they should not derive confidence from a consistent, but small sample [28]. Confidence scales could potentially mitigate the emergence of confidence-accuracy discrepancies by requiring analysts to evaluate the extent to which each piece of evidence contributes novel information to an assessment. To address the law of small numbers, standards could explicitly incorporate evidence volume and representativeness as confidence determinants, while controlling for redundancy.

In evaluating information convergence, confidence standards also fail to weigh the reliability of confirming sources against disconfirming sources [29], or how relationships between sources may unduly influence their likelihood of convergence [30]. Focusing heavily on convergence can also introduce order effects, whereby information received earlier faces fewer hurdles to being judged credible [31]. The order in which information is received should have no bearing on judgements regarding its quality, or on analytic confidence more broadly [32]. Possible solutions include the use of evidence triangulation to control for order effects, along with formalized consideration of source reliability and inter-source relationships when gauging convergence. Irwin and Mandel [31] discuss these options in the context of information evaluation but they also apply to assessments of analytic confidence.

19.5 ADDITIONAL CONFIDENCE DETERMINANTS

It is unlikely that current analytic confidence standards incorporate all relevant determinants [33], [34]. For instance, confidence levels, as traditionally expressed, fail to consider how much estimates might shift with additional information, which is often a key consideration for consumers deciding how to act on an estimate [35]. Under certain circumstances, the information content of an assessment may be less relevant to decision makers than how much that information (and the resultant estimate) may change in the future. Analytic confidence scales could incorporate a measure of "responsiveness," expressed as the probability that an estimate will change due to additional collection and analysis over a given time period (e.g., there is a 70% chance of x, but by the end of the month, there is a 50% chance that additional intelligence will increase the estimated likelihood of x to 90%) [35]. As is the case with recency, the relevant time period when assessing responsiveness varies by intelligence requirement.

In addition to responsiveness and evidence characteristics, current conceptions of analytic confidence fail to convey the level of consensus or range of reasonable opinion about a given estimate [36]. Analysts can arguably assess uncertainty more effectively when the range of plausible viewpoints is narrower, and evidence characteristics and the range of reasonable opinion vary independently [36]. In climate science, different assumptions between scientific models can lead researchers to predict significantly different outcomes using the same data. For this reason, current climate science standards incorporate model agreement/consensus as a determinant of analytic confidence (see Refs. [14], [15]). Future research could assess the most effective means of evaluating and communicating consensus among analysts. To mitigate the risk that analysts might disregard limited evidence and derive undue confidence from a high level of consensus, considerations of consensus could require simultaneous considerations of evidence volume.

Peterson [33] argues that the use of SATs, subject matter expertise, level of collaboration, task complexity, and time pressure are relevant accuracy determinants, despite being widely overlooked in current confidence standards. However, we recommend that the influence of additional determinants be empirically evaluated in the context of intelligence analysis, given that findings from the broader decision science literature may lack generalizability across tasks, and certain factors, like the use of SATs, have undergone little empirical evaluation [37], [38]. Contrary to Peterson's [33] suggestion, Tetlock [39] finds that subject matter expertise is not a significant predictor of geopolitical forecasting accuracy, although Mandel and Barnes [40] found that senior analysts had better discrimination skill than junior analysts. Meanwhile, Mandel, Karvetski, and Dhami [41] find that analysts who used the SAT Analysis of Competing Hypotheses [23] to solve a hypothesis-testing task were in fact less coherent than analysts in the no-SAT control group. These findings emphasize the importance of including determinants with a strong evidentiary basis rather than simply promulgating what appears to be a good idea. The latter approach, however, is so common in intelligence tradecraft that Mandel [38] recently coined the term, the *goodness heuristic*, to refer to the belief that what seems like a good idea ought to be treated as if it is actually a good idea and thus acted upon. For instance, as Mandel and Tetlock [42] describe, even Heuer, who was highly sensitized to the threats of confirmation bias in intelligence analysis (e.g., Ref. [43]), was resistant to testing the effectiveness of his Analysis of Competing Hypotheses SAT, which ironically Heuer had hoped would mitigate confirmation bias.

In addition to the possibility of missing determinants, current confidence standards also fail to indicate whether some determinants are more relevant than others, and whether there are interactions (positive or negative) between determinants [34]. For instance, having unrepresentative evidence may negate the confidence boost derived from a high level of convergence. We suspect that the wide range of solutions to these combination and weighting problems results in a great deal of noise being propagated by the current methods. Chang *et al.* [37] refer to the IC's apparent blind spot regarding such noise proliferation as noise neglect and outline how similar neglect pervades the IC's approach to SATs.

19.6 SCALE STRUCTURE

All of the defence and security standards examined use three or five levels on an ordinal scale to express analytic confidence. Research on standards used to communicate both information credibility / source reliability and estimative probability suggests that analysts can assign relevant values with greater precision than current scales permit [44], [45]. In other words, current standards sacrifice predictive accuracy by mandating that analysts use vague expressions to communicate confidence. Future research could assess whether analysts can reliably express confidence with greater precision, and how various degrees of precision influence consumer interpretations. Research could also evaluate the prospect of expressing analytic confidence using numerical values.

As with source reliability / information credibility, we argue that the development of a comprehensive scoring system to validly and reliably measure every relevant confidence determinant is unrealistic and overly formulaic [31]. The diversity of intelligence contexts, wide variety of relevant determinants, and potentially complicated interplay between these determinants are such that judgements about analytic confidence are irreducibly subjective. Recognizing this, an alternative to current methods would be for analysts to communicate analytic confidence numerically in the form of confidence intervals. If estimative probability is expressed as a numerical value, analytic confidence could be communicated as a probability range to qualify the estimate (for related discussion of using numerical values to communicate intelligence estimates, see chapter by Irwin and Mandel on communicating probability in intelligence in this report). For instance, an analyst could judge the probability that *x* will occur to be 75%, with a 95% confidence interval of 65 – 85 % (i.e., the analyst is 95% certain the probability lies between 65% and 85%, with 75% being the best current estimate).

Expressing confidence in this way would directly contravene the emphasis current standards place on keeping analytic confidence and estimative probability separate. However, as noted above, there is evidence that these concepts are applied interchangeably to communicate high-level intelligence judgements [6], [7], [35]. Furthermore, standards like CFINTCOM [9] and DIA Tradecraft Note 01-15 [10] emphasize the independence of confidence and probability, but blur this distinction by dictating the use of certain probability terms based on analytic confidence. CFINTCOM [9] lists *possible*, *may*, *might*, *could*, *can*, *perhaps*, *unsure*, *unknown*, *do not know*, *unable to assess*, and *undetermined* as estimative phrases to use when analytic confidence is low. Similarly, DIA Tradecraft Note 01-15 [10] proscribes the use of *unlikely*, *likely*, *improbable*, and *probably* to communicate likelihood under conditions of low analytic confidence. US JP 2-0 [11] takes the coupling further by nesting probability terms under different confidence levels. None of the extant standards correctly explain that confidence is a second-order judgement that is in fact related to probability judgements. That is, given a stated confidence level (e.g., 95% certainty), the less confident an analyst is in his or her estimate, the wider the confidence interval should be. For example, if two analysts estimate that there is a 75% chance that *x* will occur in the next 3 months, the analyst who bounds her 95% confidence interval with the values [55%, 95%] is clearly less confident than the analyst who bounds the 95% confidence interval with the values [65%, 85%].

The use of numerical confidence intervals would reduce the ambiguity inherent in interpreting verbal confidence descriptors. It would also improve analyst accountability by removing the cover provided by imprecise, non-committal ratings like *moderate*. The process of translating verbal expressions into numerical values is also shown to improve the ability of analysts to distinguish degrees of uncertainty [45], [46]. Numerical confidence intervals can be embedded in Bayesian networks, through which analysts can coherently update their evaluations as relevant determinants change or become known. Bayesian networks would also enable analysts to easily compare and pool their individual confidence judgements, thus facilitating collaboration. Communicating confidence as a probability range would also correspond to the well-established consumer preference for probabilistic information to be expressed numerically [47], [48], [49].

Friedman and Zeckhauser [50] find that intelligence consumers tend to disaggregate analytic confidence into three dimensions: reliability of available evidence, range of reasonable opinion, and responsiveness to new information. When warranted (e.g., when an estimate is particularly consequential), analysts could provide written rationales outlining how these considerations informed their confidence ranges. Analysts could also reference other determinants, such as information quality. Requiring analysts to explicitly justify their confidence assessments would further contribute to accountability, and facilitate post-mortem analysis when analysts are found to have been significantly overconfident or underconfident in an assessment. Explicit rationales would also provide consumers with useful meta-information that could be used to task further collection or analytic efforts.

We believe the combination of numerical assessment of confidence coupled with informative case-specific rationales would represent a substantial improvement over current methods. Significantly, our proposed approach would virtually negate one source of miscommunication that could lead decision makers astray. That is, when the term confidence is used in everyday language it can mean uncertainty about the judgement or claim given, but it can also be used to communicate agreement or disagreement with a judgement or claim. If a speaker makes a claim and a listener responds "I have no confidence in that claim whatsoever," most observers would agree that the listener's response implies a challenge to the first speaker's claim rather than complete uncertainty about it. Indeed, someone who claims "no confidence" in a judgement may be quite certain it is wrong. Perhaps this accounts to some degree for why many US Members of Congress weighing the DIA assessment on North Korea were apt to reject it because it was only of moderate confidence. Now it is also true that speakers are less likely to mean "I oppose this claim" when expressing low or no confidence in their own judgements. However, we still believe the ambiguity is best avoided. Communications should strive for clarity, and especially so in intelligence assessments on which consequential decision may rest.

Whether standards for communicating analytic confidence undergo a major overhaul or more incremental modifications, we emphasize the importance of grounding these changes in sound empirical research [38], [42], [51]. Too often, the IC implements measures to mitigate subjectivity and improve communication fidelity without drawing on empirical evidence, or evaluating whether these standards have the intended effect. We have observed this tendency across standards used to assess and communicate estimative probability, source reliability, information credibility, and analytic confidence. If the IC is serious about improving analytic tradecraft, and not simply promulgating standards as a means of blame avoidance following intelligence failures [52], it should make a concerted effort to exploit relevant areas of science and technology, and in particular the science of judgement and decision making.

19.7 REFERENCES

[1] Irwin, D., and Mandel, D.R. (2018). *Methods for Communicating Analytic Confidence in Intelligence to Decision Makers: An Annotated Collection*. DRDC Scientific Letter DRDC-RDDC-2018-L020. Toronto, ON: DRDC.

[2] Lowenthal, M.M. (1992). Tribal tongues: Intelligence consumers, intelligence producers. *Washington Quarterly,* 15(1):157-168.

[3] Schneider, M. (2014). The North Korean nuclear threat to the U.S. *Comparative Strategy*, 33(2):107-121.

[4] Defense Intelligence Agency. (2010). *What We Mean When We Say*. Washington DC: Defense Intelligence Agency.

[5] Lowenthal, M.M. (2016). *Intelligence: From Secrets to Policy*, (7 ed.), Washington DC: CQ Press.

[6] Office of the Director of National Intelligence. (2007). *Iran: Nuclear Intentions and Capabilities.* National Intelligence Estimate. Washington DC. Retrieved from https://www.dni.gov/files/documents/Newsroom/Reports%20and%20Pubs/20071203_release.pdf.

[7] Friedman, J.A., and Zeckhauser, R. (2012). Assessing uncertainty in intelligence. *Intelligence and National Security,* 27(6):824-847.

[8] North Atlantic Treaty Organization. (2016). *Allied Joint Doctrine for Intelligence Procedures AJP-2.1*. Brussels, Belgium.

[9] Canadian Forces Intelligence Command. (2015). *Aide-Mémoire on Intelligence Analysis Tradecraft* (v 6.0). Ottawa, ON: DND.

[10] Defense Intelligence Agency. (2015). *Expressing Analytic Certainty*. Tradecraft Note 01-15. Washington DC: Defense Intelligence Agency.

[11] United States Joint Chiefs of Staff. (2013). *Joint Publication JP 2-0, Joint Intelligence*. Washington DC. Retrieved from http://www.dtic.mil/doctrine/new_pubs/jp2_0.pdf.

[12] Public Safety Canada. (2012). *All Hazards Risk Assessment Methodology Guidelines 2012 – 2013*. Ottawa, Canada. Retrieved from https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ll-hzrds-ssssmnt/ll-hzrds-ssssmnt-eng.pdf.

[13] Office of the Director of National Intelligence. (2017). *Intelligence Community Assessment – Assessing Russian Activities and Intentions in Recent US Elections*. Retrieved from https://www.dni.gov/files/documents/ICA_2017_01.pdf.

[14] Wuebbles, D.J., Fahey, D.W., Hibbard, K.A., Dokken, D.J., Stewart, B.C., and Maycock, T.K. (Eds.). (2017). *Climate Science Special Report: Fourth National Climate Assessment, Vol. I*. Washington DC: US Global Change Research Program. Retrieved from https://science2017.globalchange.gov/downloads/CSSR2017_FullReport.pdf.

[15] United Nations, Intergovernmental Panel on Climate Change. (2014). *Fifth Assessment Report AR5 2014*. Cambridge, UK and New York, NY. Retrieved from https://www.ipcc.ch/report/ar5/.

[16] United Nations, Intergovernmental Panel on Climate Change. (2007). *Fourth Assessment Report AR4 2007*. Cambridge, UK and New York, NY. Retrieved from https://archive.ipcc.ch/publications_and_data/ar4/wg3/en/tssts-ts-2-2-decision-making-risk.html.

[17] Pachauri, R., Taniguchi, T., and Tanaka, K. (Eds.). (2000). *Guidance Papers on the Cross-Cutting Issues of the Third Assessment Report of the IPCC*. Geneva, Switzerland: World Meteorological Organization. Retrieved from https://www.ipcc.ch/pdf/supporting-material/guidance-papers-3rd-assessment.pdf.

[18] Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., van der Linden, P.J., Dai, X., Maskell, K., and Johnson, C.A. (Eds.). (2001). *Climate Change 2001: The Scientific Basis*. Retrieved from https://www.ipcc.ch/ipccreports/tar/wg1/pdf/WGI_TAR_full_report.pdf.

[19] Moss, R.H., and Schneider, S.H. (2000). Uncertainties in the IPCC TAR: Recommendations to lead authors for more consistent assessment and reporting. In: *Guidance Papers on the Cross-Cutting Issues*

of the Third Assessment Report of the IPCC, Pachauri, R., Taniguchi, T., and Tanaka, K. (Eds.), 33-51. Geneva, Switzerland: World Meteorological Organization.

[20] United Nations, Intergovernmental Panel on Climate Change. (2001). *Third Assessment Report TAR 2001*. Cambridge, UK and New York, NY. Retrieved from https://www.ipcc.ch/ipccreports/tar/.

[21] United Nations, Intergovernmental Panel on Climate Change. (2005). *Guidance Notes for Lead Authors of the IPCC Fourth Assessment Report on Addressing Uncertainties*. Cambridge, UK and New York, NY. Retrieved from http://www.ipcc-wg2.awi.de/guidancepaper/uncertainty-guidance-note.pdf.

[22] World Anti-Doping Agency. (2015). *International Standards for Testing and Investigations: Information Gathering and Intelligence Sharing Guidelines*. Retrieved from https://www.wada-ama.org/sites/default/files/resources/files/wada_guidelines-information-gathering-intelligence-sharing_final_en.pdf.

[23] Heuer, R.J., Jr. (1999). *The Psychology of Intelligence Analysis*. Washington DC: Central Intelligence Agency, Center for the Study of Intelligence.

[24] Patterson, E.S., Roth, E.M., and Woods, D.D. (1999). *Aiding the Intelligence Analyst in Situations of Data Overload: A Simulation Study of Computer-Supported Inferential Analysis under Data Overload*. Wright-Patterson AFB, OH: Air Force Research Laboratory, Human Effectiveness Directorate.

[25] Tsai, C.I., Klayman, J., and Hastie, R. (2008). Effects of amount of information on judgement accuracy and confidence. *Organizational Behavior and Human Decision Processes,* 107(2):97-105.

[26] Nickerson, R.S., and Feehrer, C.E. (1975). *Decision Making and Training: A Review of Theoretical and Empirical Studies of Decision Making and Their Implications for the Training of Decision Makers.* Technical Report NAVTRAEQUIPCEN 73-C-0128-1. Cambridge, MA: Bolt, Beranek and Newman Inc.

[27] Oskamp, S. (1982). Overconfidence in case-study judgments. In: *Judgment Under Uncertainty: Heuristics and Biases*, Kahneman, D., Slovic, P., and Tversky, A. (Eds.), 287-293. Cambridge, UK: Cambridge University Press.

[28] Thompson, J.R., Hopf-Weichel, R., and Geiselman, R.E. (1984). *The Cognitive Bases of Intelligence Analysis.* Research Report 1362. Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

[29] Capet, P., and Revault d'Allonnes, A. (2014). Information evaluation in the military domain: Doctrines, practices, and shortcomings. In: *Information Evaluation*, Capet, P., and Delavallade, T. (Eds.), 103-125. Hoboken, NJ: Wiley-ISTE.

[30] Lesot, M., Pichon, F., and Delavallade, T. (2014). Quantitative information evaluation: Modeling and experimental evaluation. In: *Information Evaluation*, Capet, P., and Delavallade, T. (Eds.), 187-228. Hoboken, NJ: Wiley-ISTE.

[31] Irwin, D., and Mandel, D.R. (2019). Improving information evaluation for intelligence production. *Intelligence and National Security*, 34(4):503-525.

[32] Tubbs, R.M., Gaeth, G.J., Levin, I.P., and Van Osdol, L.A. (1993). Order effects in belief updating with consistent and inconsistent evidence. *Journal of Behavioral Decision Making,* 6:257-269.

[33] Peterson, J.J. (2008). Appropriate factors to consider when assessing analytic confidence in intelligence analysis. Master's thesis. Erie, PA: Mercyhurst College.

[34] Wheaton, K.J. (2009). Evaluating intelligence: Answering questions asked and not. *International Journal of Intelligence and CounterIntelligence,* 22(4):614-631.

[35] Friedman, J.A., and Zeckhauser, R. (2015). Handling and mishandling estimative probability: Likelihood, confidence, and the search for Bin Laden. *Intelligence and National Security* 30 (1):77-99.

[36] Friedman, J.A., Lerner, J.S., and Zeckhauser, R. (2017). Behavioral consequences of probabilistic precision: Experimental evidence from national security professionals. *International Organization*, 71(4):803-826.

[37] Chang, W., Berdini, E., Mandel, D.R., and Tetlock, P.E. (2018). Restructuring structured analytic techniques in intelligence. *Intelligence and National Security,* 33(3):337-356.

[38] Mandel, D.R. (2019). Can decision science improve intelligence analysis? In: *Researching National Security Intelligence: National Security Intelligence: Multidisciplinary Approaches.* Coulthart, S., Landon-Murray, M., and Van Puyvelde, D. (Eds.), 117-140. Washington DC: Georgetown University Press.

[39] Tetlock, P.E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, N.J: Princeton University Press.

[40] Mandel, D.R., and Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences of the United States of America,* 111(30):10984-10989.

[41] Mandel, D.R., Karvetski, C., and Dhami, M.K. (2018). Boosting intelligence analysts' judgment accuracy: What works, what fails? *Judgment and Decision Making*, 13(6),607-621.

[42] Mandel, D.R., and Tetlock, P.E. (2018). Correcting judgment correctives in national security intelligence. *Frontiers in Psychology*, 9:2640.

[43] Heuer, R.J., Jr. (1987). Nosenko: Five paths to judgment. *Studies in Intelligence,* 31(3):71-101.

[44] Samet, M.G. (1975). *Subjective Interpretation of Reliability and Accuracy Sales for Evaluating Military Intelligence*. Technical Paper 260. Arlington, VA: US Army Research Institute for Behavioral and Social Sciences.

[45] Friedman, J.A., Baker, J.D., Mellers, B.A., Tetlock, P.E., and Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly,* 62(2):410-422.

[46] Tetlock, P.E., and Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. New York, NY: Crown Publishing Group.

[47] Murphy, A.H., Lichtenstein, S., Fischhoff, B., and Winkler, R.L. (1980). Misinterpretation of precipitation probability forecasts. *Bulletin of the American Meteorological Society,* 6:695-701.

[48] Brun, W., and Teigen, K.H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes,* 41(3):390-404.

[49] Wallsten, T.S., Budescu, D.V., Zwick, R., and Kemp, S.M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society,* 31:135-138.

[50] Friedman, J.A., and Zeckhauser, R. (2018). Analytic confidence in political decision making: Experimental evidence from national security professionals. *Political Psychology,* 39(5):1069-1087.

[51] Dhami, M.K., Mandel, D.R., Mellers, B.A., and Tetlock, P.E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science,* 10(6):753-757.

[52] Tetlock, P.E., and Mellers, B.A. (2011). Intelligent management of intelligence agencies: Beyond accountability ping-pong. *American Psychologist,* 66(6):542-554.



Chapter 20 – COMMUNICATING ANALYSIS IN THE DIGITAL ERA AND THE COMMUNICATION OF UNCERTAINTY IN COMMERCIAL OPEN-SOURCE INTELLIGENCE[1]

Rubén Arcos
Rey Juan Carlos University
SPAIN

20.1 THE CHANGING COMMUNICATIONS ENVIRONMENT

In the intelligence studies literature, academic discussion, and even professional practice, the communication of analytic products to government decision makers as a step in the intelligence process has traditionally received much less attention than intelligence collection and analysis. Without the effective communication of intelligence to policymakers, however, all the previous efforts in collection and analytic production are futile. This process involves activities that require the acquisition of specific competencies to be conducted successfully.

The communication domain is driving huge transformations around the world in virtually all industries, changing the ways we interact, build, and conduct relations with others. It is difficult to imagine a field of human experience that has not been transformed by digital information and communication technologies. The editorial staffs of newspapers, news agencies, and other traditional media all over the world have witnessed and are experiencing major transformations driven by technological innovations while they struggle to adapt the processes of producing and delivering news and information of relevance to their readers.

Multimedia communication among individuals is nowadays usual in both professional and private spheres. Multimedia can be defined as "any combination of text, graphics, video, audio, and animation in a distributable format that consumers can interact with using a digital device" [2].

Newspapers provide highly attractive interactive infographics in their digital editions, presenting data and information in a digestible format. Videos increasingly are being introduced in digital publications. People are growing increasingly familiar with digital technologies and are both consumers and producers of digital photography, video, and blogging and micro-blogging platforms. They are consumers of information but are also learning how to produce content for others.

Coinciding with this ongoing transformation, people are getting used to new ways of consuming information and interacting with that information. They are assuming a more active role in the process. Consequently, decision makers in government and industry are demanding – and will increasingly demand – analysis and intelligence products adapted to this new era of digital communication. As the threshold of usefulness for new information decreases, the consumer of intelligence analysis is expecting shorter timeframes for receiving intelligence products (see Figure 20-1).

20.2 THE INTELLIGENCE USER EXPERIENCE (UX)

As reported in the *Washington Post*, the iPad has already been used to disseminate intelligence to US President Obama, "allowing analysts to add video and audio clips and interactive graphics" [3]. The

[1] Most of the material in this chapter was previously published in Ref. [1], reproduced with permission of Palgrave Macmillan. This chapter is based on the paper "Producing and Consuming Intelligence Products in the Digital Era: The Need for Multimedia Communications," prepared by the author for presentation at the panel Reinventing Intelligence Production for the 21st Century, International Studies Association, Toronto, Canada, 29 March 2014.

article quotes Shawn Turner, director of Public Affairs for the Office of the Director of National Intelligence, describing tablets as a proper and secure alternative way to provide intelligence and likely to be used more frequently in the future to represent multimedia information in the PDB[2].

Although it is the responsibility of each nation's intelligence community to establish its own communication standards for analytic products, the very purpose of the intelligence is to deliver the best possible analyses in a timely and usable way to facilitate the decision making of the policymaker or consumer. Intelligence products are useful only if they provide information and analytic insights on time to the decision maker, reducing their uncertainty and the tension that it produces while facilitating decisions and posterior actions. Independent of whether the intelligence product is conceived to be delivered in a printed or in a digital and interactive format, the product needs to be delivered to the user in a timely fashion.

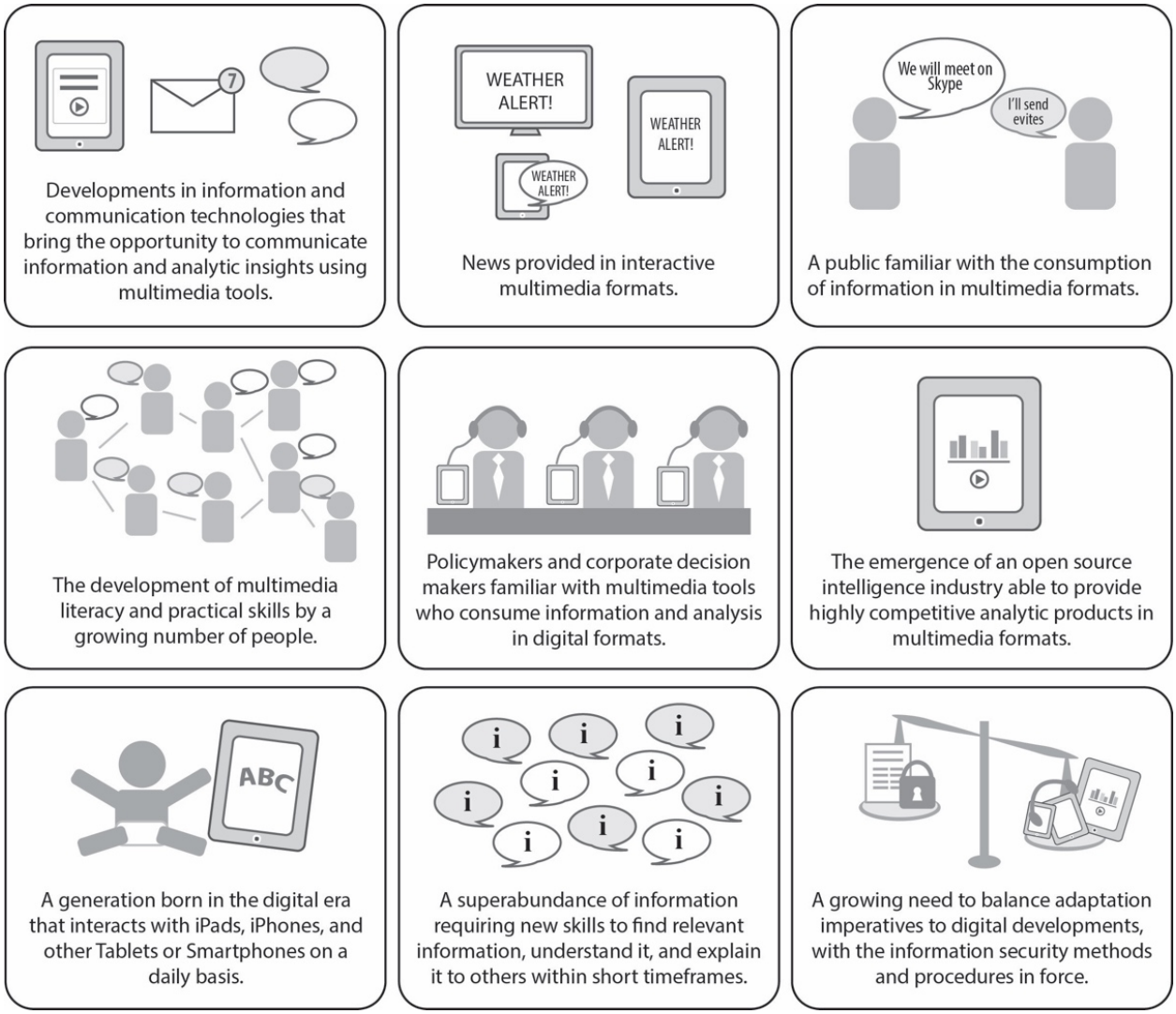


Figure 20-1: Characteristics of the Emerging Communications Environment.

[2] An official photo (dated January 31, 2012) of President Barack Obama using a tablet computer while receiving the Presidential Daily Briefing can be seen at http://www.whitehouse.gov/photos-and-video/photo/2012/01/president-barack-obama-receives-presidential-daily-briefing.

In principle, immediacy is one of the advantages of digital communication compared to printed media. In the absence of the necessary skills to design information (or intelligence), combine and integrate media in a meaningful way, and anticipate sequences of interaction by the user, however, the results can be counterproductive. Although the traditional principles of analytic writing remain essential, a core difference exists between textual reports and digital intelligence products:

> *"All traditional text, whether in printed form or in computer files, is sequential, meaning that there is a single linear sequence defining the order in which the text is to be read […] Hypertext is nonsequential; there is no single order that determines the sequence in which the text is to be read […] Hypertext presents several different options to the readers, and the individual reader determines which of them to follow at the time of reading the text. This means that the author of the text has set up a number of alternatives for readers to explore rather than a single stream of information […] hypertext consist of interlinked pieces of text (or other information)."*[4]

A "digital turn" in the field of intelligence communication needs take into consideration the fields of information design and interaction design, as well at the concept of User Experience (UX). All of them provide a framework that sets the stage for digital communication and design for interactive media that can be useful for intelligence communication in the 21st century.

As noted by Hartson and Pyla, the concepts of UX and design do not necessarily entail high-tech artifacts; technology is rather a design context [5]. Similarly, the concept of usability is critical. The user experience can be influenced by perceptions of the producing organization and past experiences. According to Nielsen, usefulness, defined as the capability of a system to be used to achieve a goal, can be broken down into two categories: usability and utility. Specifically:

> *"Utility is the question of whether the functionality of the system can do what is needed, and usability is the question of how well users can use the functionality."* [6]

The concept of usability applies to all aspects related to the systems with which we interact. It consists of five attributes: learnability (easy to learn), efficiency (efficient to use), memorability (easy to remember), errors (low error rate), and satisfaction (subjectively pleasant to use) [6].

Accordingly, intelligence products should be designed by taking usability and its components into account. Table 20-1 applies these usability attributes to the interaction of the intelligence consumer/user with analytic products.

In the field of interaction design, UX integrates the concepts of utility and usability but adds other components. UX is an expansion of the concept of usability design, entailing also "social and cultural interaction, value-sensitive design, and emotional impact – how the interaction experience includes joy of use, fun, and aesthetics" [5]. UX is defined as:

> *"The totality of the effect or effects felt by a user as a result of interaction with, and the usage context of, a system, device, or product, including the influence of usability, usefulness, and emotional impact during interaction, and savoring the memory after interaction. Interaction with is broad and embraces seeing, touching, and thinking about the system or product, including admiring it and its presentation before any physical interaction."* [5]

Adopting a user-focused approach when producing analyses is critical in order to be relevant for intelligence consumers. The concept of Intelligence UX highlights this necessity by considering not only the most fundamental aspect of the utility of intelligence products, but also how it is related with their usability through a proper interaction design (of the user with the product). The design should take into account that satisfactory use (joy of use) by the client will have an impact in his/her willingness to use the system for making decisions.

Table 20-1: Application of Usability to Traditional Analytic Products.

Usability Attributes	Meaning	Application to Analytic Products
Learnability	Easy to learn	The structural organization (inverted pyramid approach) of the product facilitates interaction by the user. Degrees of uncertainty and the quality of sourcing are expressed using an easy system of words/numbers.
Efficiency	Easy to use	The paper presents a clear picture, addressing the "so what" and putting the bottom line up front, with key judgements and implications highlighted.
Memorability	Easy to remember	Layout template is easy to remember (title, headings, bottom line, key judgements) and the paper tells a compelling story.
Errors	Low error rate	Avoidance of misspellings, grammatical errors, unfounded assumptions, and poor logic.
Satisfaction	Pleasantly used	Preference for one system (analytic product) over others, visually effective, attractive layout, and good use of graphics.

It seems plausible that products that provide a satisfactory UX are more likely to impact decisions. In a world of information and cognitive overload, the analyst has to struggle to capture the attention of the intelligence client. It is not enough, although desirable and certainly the most important for the intelligence service mission, to collect the best possible information and provide the best possible analysis. Much more than in the past, intelligence analyses competitiveness is now affected by the manner in which an insightful analysis and strategic information is conveyed to consumers.

Although it may seem obvious, it is important to highlight that in an environment of multi-touch screens where digital communication is the norm, organizations that only deliver printed products will be perceived as an exception. The expectation will be that high-quality products must include digital input. In addition, narratives from entertainment industry and images of high-tech intelligence services affect the cultural environment influencing the policymaker. In the corporate world, it will become unthinkable to provide competitive and market intelligence deliverables that are devoid of graphics and digital media.

20.3 MULTIMEDIA COMPETENCIES AND THE INTELLIGENCE ANALYST

The concepts of digital natives [7], net generation (1977 – 1997) [8], or generation C [9], among others, stress the implications derived from the impacts produced by the technological changes in the generations that have grown in the Digital Era. According to Palfrey and Gasser, digital natives "were all born after 1980, when social digital technologies such as Usenet and bulletin board systems came online? They all have access to networked digital technologies. And they all have the skills to use those technologies" [10].

Although timeframes vary depending on sources, most authors agree that digital technologies have sociological, economic, psychological, political, and cultural consequences:

> *"They are Generation C – connected, communicating, content-centric, computerized, community-oriented, always clicking. As a rule, they were born after 1990 and lived their adolescent years after 2000. In the developed world, Generation C encompasses everyone in this age group; in the BRIC countries, they are primarily urban and suburban. By 2020, they will make up 40 percent of the population in the United States, Europe, and the BRIC countries, and 10 percent in the rest of the world – and by then, they will constitute the largest group of consumers worldwide."* [9]

Thus, the trend is a massive use of digital technologies both for the next and current generations. However, being able to use digital technologies and consume products in digital formats is quite different than being skilled at producing contents in digital formats. Education for providing digital literacy and training digital communication skills is required for being able to use digital technologies from the perspective of a producer. The author's experience teaching multimedia communication courses to students of journalism for several years as well as designing and conducting with colleagues the *Multimedia Intelligence Product* simulation exercise as part of a Master's Degree program in intelligence analysis for four years shows that being a user of digital communication devices and information does not equate with being a skillful producer.[3] In our classes, some Generation C users experience paralysis, spurred by their fear of technology, and are unable to perform clear step-by-step instructions when asked to build rather than to use digital devices. Getting familiar with a specific multimedia communication-related language and tools can pose additional challenges, particularly if the task involves computer programming.

Overcoming the initial frustration that results from not being able to successfully complete a task, understanding the reasons behind recurrent unsuccessful attempts, as well as activities such as successfully publishing for the first time a Website or updating the version of a Content Management System are milestones that appear like rites of passage. The practice of digital journalism requires professionals with specific knowledge and skills. Higher education programs and continuing learning courses provide education and training in digital communication competencies. A digital turn in intelligence analysis communication requires practically the same competencies of digital journalism, and specific training that stresses the existing differences in both practices.

Usability attributes also apply to the producer. Utility and usability are mandatory. The usefulness of an intelligence product from the perspective of the producer depends on how much additional value the system provides for better achieving the organization's mission and creating informative value for facilitating the decision-making process. Insightfulness and timeliness are key. The system has to be usable and facilitate timely and satisfactory experiences. Timeliness is a priority. Back-end technology needs to be usable and to facilitate an optimal workflow. The number and complexity of tasks to perform by the analysts in order to convey the best possible analyses in multimedia formats has to be not higher than the ones performed when writing analytic papers in text-only format. A mockup of a graphical user interface model for guiding analysts through the process of writing the intelligence reports, including help tooltips and basic reminders on the principles of analytic writing, was presented by the author at the SCIP European Summit 2013 [12].

20.4 WIREFRAMING AN INTELLIGENCE REPORT

How does a multimedia report look? The answer to the question depends on the nature of the intelligence product. The conceptualization and design of the products will depend on whether the analysis is of a basic, current, estimative, or indications and warning nature. It will also depend on the mission and culture of the organization and should be informed by consumer feedback.

Based upon the structure of an unclassified US National Intelligence Council report, Figure 20-2 sketches the information structure, labeling, and interaction of a hypothetical digital intelligence report [13]. This wireframe of the multimedia report is similar to one proposed to students as a template for running the simulation exercise previously cited. The participants produce these multimedia reports as part of their training.

Specific products of other intelligence communities like unclassified UK Joint Intelligence Committee (JIC) assessments can also inspire the structure of the reports. The static representation of the interactive product involves navigation through the following tabs and buttons:

[3] The simulation has been published in a ready-to-run format [11]. Participants have widely benefited from the expertise of professors Sergio Álvarez and Manuel Gértrudix in the field of digital communication.

- Scope Note;
- Bottom Line;
- Key Judgements;
- Background;
- Assumptions;
- Video Briefing;
- Scenarios;
- Sourcing and References (including source evaluation); and
- Export as E-paper (an embedded intelligence report that can be exported as in portable document format).

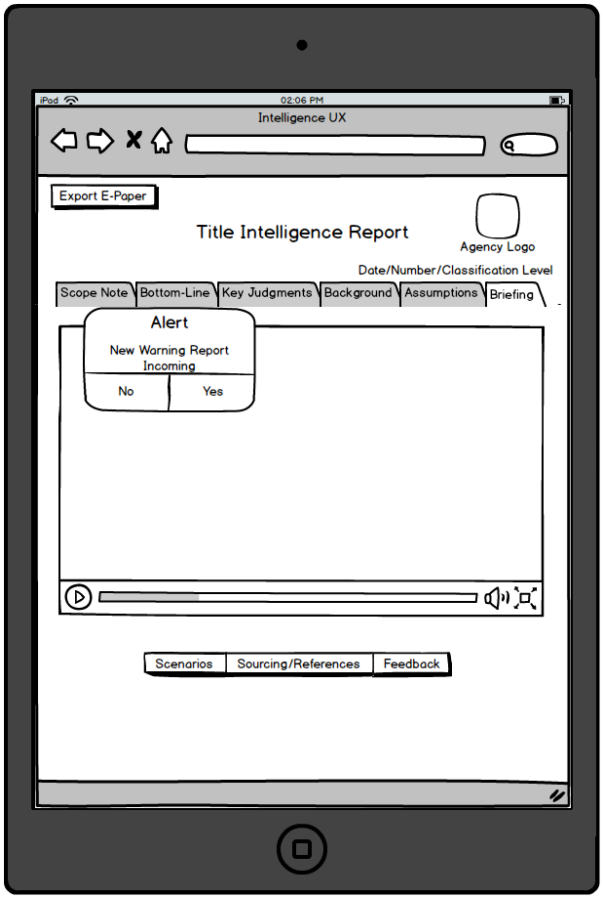


Figure 20-2: Wireframe of a Multimedia Report Created with Balsamiq Mockups [1].

Given the non-sequential nature of digital products, it is important to establish a product map and anticipate scenarios and paths for the user. Labelling should be consistent with the content and support easy navigation. An opportunity provided by digital communication is that oversight bodies or internal reviewers can introduce additional requirements for evaluating the correctness of the analysis and make more transparent the analytic process allowing its evaluation. Digital products can help in tracking analytic bias.

A variety of multimedia elements can assist in providing valuable intelligence products to consumers. The question is not how many of them to include but which are most appropriate. Interactive infographics, maps,

timelines, static image sliders, audio, customizable and interactive charting, diagrams, network visualization, and explanatory videos can be integrated in the report. Many of these tools and solutions can already be found in the commercial market. The business community uses several well-known examples of competitive intelligence proprietary software.

A problem that usually surfaces with the use of this software is the lack of an appropriate information design and internal organization of the reports. In other words, analysis has to be provided through a carefully structured presentation of the key judgements and findings and supported by a solid argumentation. It is not just facts. It is about extracting and explaining meanings and implications, as well as anticipating developments.

Garret's conceptual framework for designing user experience in the case of Websites provides a clear model that can be useful for conceptualizing multimedia intelligence reports [14]. According to Garret's framework, five planes affect the composition of a Web site: strategy, scope, structure, skeleton, and surface. Each plane is dependent on the planes below it (see Figure 20-3).

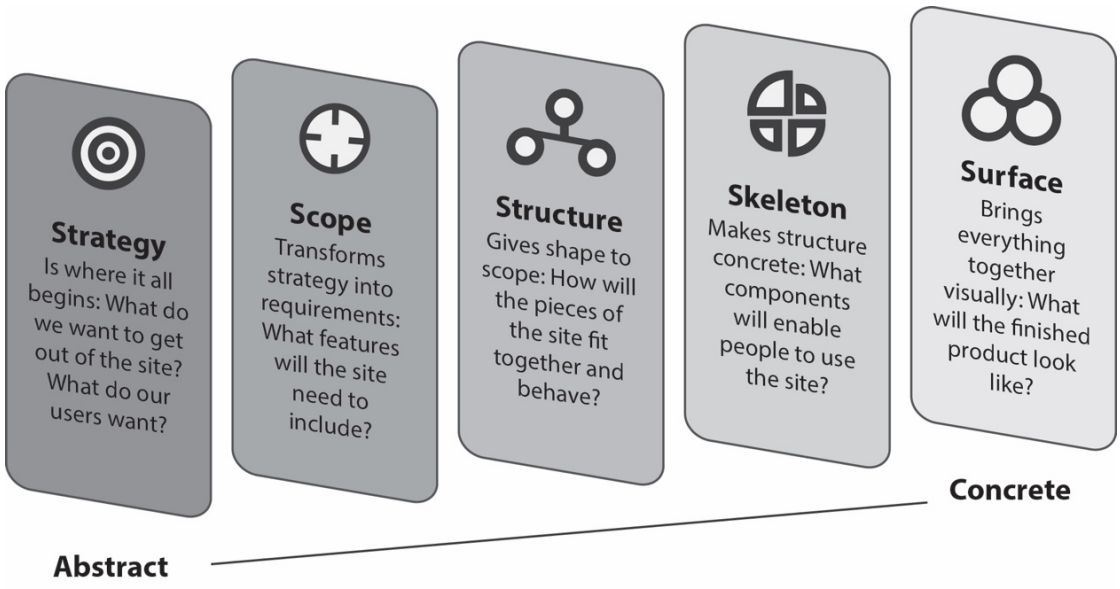


Figure 20-3: The Five Planes of the User Experience.

- The **strategy plane** considers the expectations of the user about the product, as well as the organization objectives in relation to the product and those needs.
- The **scope plane** translates the strategy plane into specific requirements "for what content and functionality the product will offer to the users" [14]. Content requirements refer both to text and multimedia elements like the ones above mentioned, but it is necessary to remember that the purpose of the content is the primary concern. Format should not make us forget the informative and cognitive value of the analysis.
- Garret links the **structure plane**, which brings concretion to the product, to interaction design and information architecture. The first has to do with how the user will interact with the systems and how the system will react to the user behavior. Information architecture is concerned with how contents are structured for facilitating understanding and use ranging from hierarchical to free exploration organic sites.
- The **skeleton plane** depends on navigation design, information design, and the design of the interface. Wireframes connect the three of them in a schematic design. According to Treder,

wireframes are low fidelity depictions of a design showing core groups of content, the structure of information, and basic visualizations of actions between users and the interface, a kind of backbone of the design, which is consistent with Garret's concept of the skeleton plane [15].

- The **surface plane** deals with concreting the visual design and what Garret names sensory design. This includes color palettes, typography, and branding elements like the logo of the organization [14].

20.5 STRUCTURED BRIEFINGS AND FACE-TO-FACE INTERACTION

Multimedia intelligence products and digital communications should not be considered as substitutes for face-to-face interaction with the decision maker. Rather, multimedia and digital communications are aids supporting the explanatory function of intelligence analysis. Interpersonal interaction is key to build trust and emotional involvement. Technology in this context is conceived to support understanding and communication, not to increase the distance with the client. For instance, if the client is too senior or too busy, a video briefing by an experienced analyst is better than a no-briefing scenario.

Face-to-face presentations require specific skills and preparation. Deep knowledge and experience build trust but do not guarantee a persuasive and compelling discourse or a superior speaking performance.

Immersive communication and augmented reality open new ways to interact with the customers and provide briefings. Again, bringing structure to the content under a proper timeframe is key. Digital communication provides opportunities to increase the number of presentations. Digital communication and augmented reality will compel analysts and their managers to rethink the portfolio of intelligence deliverables and presentations aimed at supporting the decisions of the C-Suite. In the case of intelligence services and classified information, digital communications can be challenging to information security policies and standards. However, the need to adapt–sooner or later–hardcopy, narrative analytic products to the digital era will be necessary to remain competitive. It is the norm to find high levels of resistance to change within intelligence organizations with a toughly rooted secrecy culture. In these cases, it is important to remind them–as observed by Berkowitz and Goodman–that secrecy should be considered a tool of the trade; one that imposes costs and that should be managed intelligently [16].

20.6 THE COMMUNICATION OF ANALYSIS AND UNCERTAINTY IN COMMERCIAL OPEN-SOURCE INTELLIGENCE PUBLICATIONS: THE CASE OF *JANE'S INTELLIGENCE REVIEW*

The analytic outputs of intelligence services are not the only intelligence feeds used to support decision making on security and defence issues. The providers of commercial open-source products can also inform decisions and be used in the intelligence process. Open-source products convey uncertainty and risks following guidelines that can be similar or not to the systems used by intelligence services. Thus, there is a need for knowing and understanding how commercial open-source providers produce analyses and assessments, and more specifically how they structure their products and communicate degrees of uncertainty and risk.

Jane's Intelligence Review (JIR) provides a relevant case study with this regard. This publication describes itself as delivering "unbiased intelligence and analysis on the most critical and decisive international security issues and country risks" [17]. JIR articles are published online and once monthly in print and digital formats, covering international security issues in different sections and specific kind of articles ranging from summaries of news developments produced in-house to longer feature pieces of around 3,500 words open to freelance contributing experts. The structure of the articles varies according to the section and the article length. However, the articles tend to follow a "standard model" that has similarities with the UK JIC assessments and

EU Intelligence and Situation Centre (INTCEN) assessments. This model includes an internal structure consisting on the following sections: title; subtitle; key points; body of the article, including background, analysis and assessment; and outlook. The articles may include relevant pictures, maps, cases and diagrams.

How does JIR convey uncertainty in analytic judgements? According to the Jane's May 2017 Guidelines for freelance contributors, JIR tries to adhere to the US Intelligence Community Directive (ICD) 203 of 2 January 2015 for the expressions of probability in assessments. It is important to consider this standard for the expression of probability when consuming or using the analytic judgements contained in Jane's open-source pieces. See Table 20-2.

Table 20-2: Expressions of Probability in *Jane's Intelligence Review* According to the US ICD-203 of 2 January 2015 [20].

Almost no chance	1 – 5 %
Very unlikely	5 – 20 %
Unlikely	20 – 45 %
Roughly even chance	45 – 55 %
Likely	55 – 80 %
Highly likely	80 – 95 %
Almost certain	95 – 99 %

The adoption of a standardized lexicon for communicating uncertainty with associated numerical probabilities by commercial open-source intelligence providers has the obvious advantage of bringing transparency and using the same system as the one employed by intelligence institutions, and can be useful for reducing the ambiguity and randomness in the use of language. However, the academic research on the communication of uncertainty in intelligence analysis has shown that most existing lexicons "are not based on empirical evidence, and their effectiveness has not been empirically tested" [18]. In particular, a study by Ho, Budescu, Dhami, and Mandel found that "the abbreviated NIC lexicon achieved a mean consistency rate similar to those found using our methods, but it failed to cover the entire range of interest. In particular, the NIC lexicon neglects very high and very low probabilities whose reporting to decision makers can be highly consequential in intelligence analysis" [18].

On the communication of risk, *Jane's Intelligence Review* "scores risk numerically in dynamic indices that are comparable across countries ranging from 0.1 to 10, with steps 0.1 on a logarithmic scale". [4] According to JIR's editor, Robert Munks, "the country risk ratings are: Low (0.1 – 0.7); Moderate (0.8 – 1.5); Elevated (1.6 – 2.3); High (2.4 – 3.1); Very High (3.2 – 4.3); Severe (4.4 – 6.4); Extreme (6.5 – 10.0)" [19]. These ratings are associated with colors as well and then used in the Scenarios section together with radar charts showing the risk environment in specific countries as shown in Figure 20-4. The same system for communicating risk is employed in the Outlook section of this publication that provides a world map with "security flashpoints".

While maps and images illustrating the open-source pieces have been employed in the past, in the recent years JIR has increased the number of infographics, diagramming, imagery analysis, and visual explanatory

[4] On the country risk ratings, Robert Munks explains that "this is a logarithmic scale, so the variance tends to be most heavily concentrated in the lower bands, and very few countries emerge with risk categories in 'severe' and virtually none in 'extreme' (which would be pretty much a nuclear war scenario)". Personal exchange by the author with Jane's Intelligence Review editor, Robert Munks, May 17, 2017. See also the Scenarios section of Jane's Intelligence Review, in, for example Ref. [19].

materials accompanying the articles. The magazine is now published as well in a digital edition to be read using a web browser or though the IHS Defence Magazines app for iOS devices, although the degree of interactivity of the user with the digital product –for example, the interaction with charts and other visual representations – has room for improvement.



Figure 20-4: Jane's Country Risk Scale[5].

20.7 THE VISUAL COMMUNICATION OF UNCERTAINTY

For the reasons discussed in the preceding sections it seems plausible that visual methods for representing uncertainty will be introduced in intelligence flagship publications and commercial open-source intelligence products. However, existing academic research on the visualization of uncertainty has concluded that:

> *"A wide variety of strategies has been proposed for uncertainty visualization, and there is a growing body of empirical research that is providing insights concerning which methods are effective in different use contexts and for which types of tasks. Still, we have only scratched the surface of the problem. For example, we cannot yet say definitively whether decisions are better if uncertainty is visualized or suppressed, or under what conditions they are better; nor do we understand the impact of uncertainty visualization on the process of analysis or decision making."* [21] (See also Ref. [22].)

One of the various challenges identified in the relevant literature consists of "assessing the usability and utility of uncertainty capture, representation, and interaction methods and tools […] we need progress toward two additional challenges: new empirical methods are needed to study the use of highly interactive display forms; and new empirical methods are needed for studying the role of visual representation as an input to strategies for addressing ill-structured problems" [23].

Among the visual methods proposed for communicating uncertainty, different graphic variables have been proposed such as: color saturation (varying from pure hues for very certain to unsaturated colors for high degrees of uncertainty) and symbol focus (from "in focus" for depicting higher degrees of certainty to "out of focus" for uncertainty representation) [23] (see also Ref. [24]). According to MacEachren (1992), symbol focus can be manipulated for depicting uncertainty through:

1) Symbol contour "fuzziness": varying from sharp narrow lines (certainty) to broad fuzzy lines (uncertainty);
2) Symbol fill clarity;
3) Transparency or fog: where uncertainty is communicated through the thickness of the fog from more transparent to nebulous; and
4) Image resolution [24].

[5] Scale used in the permanent Security Flashpoint section of the digital and printed magazine.

Location has been proposed as well as a visual variable for communicating uncertainty where further from center represents a higher degree of uncertainty [25].

20.8 CONCLUSION

Digital communication opens new ways for communicating analysis and assessments to intelligence users and can potentially improve the experience of decision makers in the consumption of intelligence products. The introduction of visual, multimedia, and interactive elements however needs to be based on the evidence provided by findings of research studies on the effectiveness of methods and tools used. The concepts of user experience and usability, with its constituent attributes, should be considered when designing digital intelligence publications and producing intelligence analyses that introduce multimedia and interactive elements.

20.9 REFERENCES

[1] Arcos, R. (2015). Communicating analysis in a digital era. In: *Intelligence Communication in the Digital Era: Transforming Security, Defence and Business*, Arcos, R., and Pherson, R.H. (Eds.), 10-23. London, UK: Palgrave Pivot. https://link.springer.com/chapter/10.1057/9781137523792_2. Reproduced with permission of Palgrave Macmillan.

[2] Costello, V., Youngblood, S.A., and Youngblood, N.E. (2012). *Multimedia Foundations: Core Concepts for Digital Design*, 12. Waltham, MA: Elsevier.

[3] Miller, G. (2012). Oval Office iPad: President's daily intelligence brief goes high-tech. Checkpoint Washington. Retrieved from http://www.washingtonpost.com/blogs/checkpoint-washington/post/oval-office-ipad-presidents-daily-intelligence-brief-goes-high-tech/2012/04/12/gIQAVaLEDT_blog.html.

[4] Nielsen, J. (1995). *Multimedia and Hypertext: The Internet and Beyond,* 1-2. Mountain View, CA: Morgan Kaufmann.

[5] Hartson, R., and Pyla, P.A. (2012). *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*. Waltham, MA: Morgan Kaufmann.

[6] Nielsen, J. (1993). *Usability Engineering*, 25. Mountain View, CA: Morgan Kaufmann.

[7] Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon,* 9(5). Retrieved from http://marcprensky.com/articles-in-publications/.

[8] Tapscott, D. (2009). *Grown Up Digital: How the Next Generation Is Changing Your World*. New York, NY: McGraw-Hill.

[9] Friederich, R., Peterson, M., Koster, A., and Blum, S. (2010). The rise of Generation C and implications for the world of 2020. Booz & Company. Retrieved from http://www.booz.com/media/file/Rise_Of_Generation_C.pdf.

[10] Palfrey, J., and Gasser, U. (2008). *Born Digital: Understanding the First Generation of Digital Natives*. New York, NY: Basic Books.

[11] Arcos, R., Gértrudix, M., Prieto, J.I. (2014). Multimedia intelligence products: Experiencing the intelligence production process and adding layers of information to intelligence reports. In: *The Art of Intelligence: Simulations, Exercises, and Games*, Lahneman, W.J., and Arcos, R. (Eds.), 239-260. Lanham, MD: Rowman & Littlefield Publishers.

[12] Arcos, R., and Gértrudix, M. (2013). Apply multimedia technology for intelligence reporting. Presentation delivered at the European SCIP Summit 2013, Rome, Italy, 5 – 7 November.

[13] Global Water Security. *Intelligence Community Assessment*. ICA 2012-08, Retrieved from http://www.dni.gov/files/documents/Special%20Report_ICA%20Global%20Water%20Security.pdf (2 February 2012).

[14] Garrett, J.J. (2011). *The Elements of User Experience: User-Centered Design for the Web and Beyond,* 2nd ed. Berkeley, CA: New Riders Press.

[15] Treder, M. (2013). *UX Design for Startups*. Retrieved from www.uxpin.com.

[16] Berkowitz, B.D., and Goodman, A.E. (2000). *Best Truth: Intelligence in the Information Age*, 160. New Haven, CT: Yale University Press.

[17] Jane's Group (2020) Jane's Online Shop. https://shop.janes.com/Magazines/Jane-s-Intelligence-Review/. (19 March 2020)

[18] Ho, E.H., Budescu, D.V., Dhami, M.K., and Mandel, D.R. (2015). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science & Policy,* 1(2):51.

[19] Machenheimer, S., and Aboo, S. (2017). Age of uncertainty: factional politics threaten Zimbabwe's stability. IHS Jane's Intelligence Review 29(2): 18-21.

[20] IHS Markit (2017). Writing Articles for Jane's Intelligence Review: Guidelines for Freelance Contributors, revised 10 May 2017 (Rev6).

[21] MacEachren, A.M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., and Hetzler, E. (2005). Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):155.

[22] Kinkeldey, C., MacEachren, A.M., Riveiro, M., and Schiewe, J. (2017). Evaluating the effect of visually represented geodata uncertainty on decision making: Systematic review, lessons learned, and recommendations. *Cartography and Geographic Information Science,* 44(1):18.

[23] MacEachren, A.M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., and Hetzler, E. (2005). Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science,* 32(3):139-160.

[24] MacEachren, A.M. (1992). Visualizing uncertain information. *Cartographic Perspective,* 13(Fall 1992): 10-19.

[25] MacEachren, A.M., Roth, R.E., O'Brien, J., Li, B., Swingley, D., and Gahegan, M. (2012). Visual semiotics and uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics,* 18(12):2496-2505.

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's References	3. Further Reference	4. Security Classification of Document
	STO-TR-SAS-114 AC/323(SAS-114)TP/928	ISBN 978-92-837-2253-3	PUBLIC RELEASE

5. Originator

Science and Technology Organization
North Atlantic Treaty Organization
BP 25, F-92201 Neuilly-sur-Seine Cedex, France

6. Title

Assessment and Communication of Uncertainty in Intelligence to Support Decision-Making

7. Presented at/Sponsored by

Final Report of Research Task Group SAS-114

8. Author(s)/Editor(s)	9. Date
David R. Mandel	June 2020

10. Author's/Editor's Address	11. Pages
Department of National Defence PO Box 2000 1133 Sheppard Ave West Toronto ON M3J 2C9 Canada	384

12. Distribution Statement

There are no restrictions on the distribution of this document. Information about the availability of this and other STO unclassified publications is given on the back cover.

13. Keywords/Descriptors

Accuracy; Decision-making; Forecasting; Information credibility; Information evaluation; Intelligence analysis; Intelligence assessment; Intelligence production; Judgment; Probability; Reasoning; Source reliability; Structured analytic techniques; Uncertainty

14. Abstract

The primary objective of the SAS-114 Research Task Group was to investigate methods of improving intelligence assessments and communicating the uncertainties surrounding such assessments clearly to decision-makers. For this undertaking, SAS-114 members collected and evaluated a wide range of uncertainty communication standards currently used in defence and security as well as in other domains including law enforcement, climate science, and medicine. This analytic task addressed two broad problem areas: (a) the promulgation of multiple, inconsistent uncertainty communication standards within and across nations and (b) the use of standards that are either conceptually flawed or poorly suited to the specific context of application. SAS-114 members also conducted original research and analysis on topics including monitoring forecasting skill in intelligence; evaluating structured analytic techniques for improving intelligence assessment; and the incorporation of evidence-based techniques into analytical tradecraft and training. A key recommendation of SAS-114 is that intelligence organizations should leverage judgment-and-decision science to verify and improve the quality of their analytic processes and, by extension, intelligence products. By systematically drawing on relevant scientific evidence, intelligence organizations could improve their accuracy, rigour, and communication fidelity, which in turn should better support sound decision-making and interoperability within NATO.



NORTH ATLANTIC TREATY ORGANIZATION

BP 25
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@cso.nato.int

SCIENCE AND TECHNOLOGY ORGANIZATION

DIFFUSION DES PUBLICATIONS
STO NON CLASSIFIEES

Les publications de l'AGARD, de la RTO et de la STO peuvent parfois être obtenues auprès des centres nationaux de distribution indiqués ci-dessous. Si vous souhaitez recevoir toutes les publications de la STO, ou simplement celles qui concernent certains Panels, vous pouvez demander d'être inclus soit à titre personnel, soit au nom de votre organisation, sur la liste d'envoi.

Les publications de la STO, de la RTO et de l'AGARD sont également en vente auprès des agences de vente indiquées ci-dessous.

Les demandes de documents STO, RTO ou AGARD doivent comporter la dénomination « STO », « RTO » ou « AGARD » selon le cas, suivi du numéro de série. Des informations analogues, telles que le titre est la date de publication sont souhaitables.

Si vous souhaitez recevoir une notification électronique de la disponibilité des rapports de la STO au fur et à mesure de leur publication, vous pouvez consulter notre site Web (http://www.sto.nato.int/) et vous abonner à ce service.

CENTRES DE DIFFUSION NATIONAUX

ALLEMAGNE
Streitkräfteamt / Abteilung III
Fachinformationszentrum der Bundeswehr (FIZBw)
Gorch-Fock-Straße 7, D-53229 Bonn

BELGIQUE
Royal High Institute for Defence – KHID/IRSD/RHID
Management of Scientific & Technological Research
for Defence, National STO Coordinator
Royal Military Academy – Campus Renaissance
Renaissancelaan 30, 1000 Bruxelles

BULGARIE
Ministry of Defence
Defence Institute "Prof. Tsvetan Lazarov"
"Tsvetan Lazarov" bul no.2
1592 Sofia

CANADA
DGSIST 2
Recherche et développement pour la défense Canada
60 Moodie Drive (7N-1-F20)
Ottawa, Ontario K1A 0K2

DANEMARK
Danish Acquisition and Logistics Organization
(DALO)
Lautrupbjerg 1-5
2750 Ballerup

ESPAGNE
Área de Cooperación Internacional en I+D
SDGPLATIN (DGAM)
C/ Arturo Soria 289
28033 Madrid

ESTONIE
Estonian National Defence College
Centre for Applied Research
Riia str 12
Tartu 51013

ETATS-UNIS
Defense Technical Information Center
8725 John J. Kingman Road
Fort Belvoir, VA 22060-6218

FRANCE
O.N.E.R.A. (ISP)
29, Avenue de la Division Leclerc
BP 72
92322 Châtillon Cedex

GRECE (Correspondant)
Defence Industry & Research General
Directorate, Research Directorate
Fakinos Base Camp, S.T.G. 1020
Holargos, Athens

HONGRIE
Hungarian Ministry of Defence
Development and Logistics Agency
P.O.B. 25
H-1885 Budapest

ITALIE
Ten Col Renato NARO
Capo servizio Gestione della Conoscenza
F. Baracca Military Airport "Comparto A"
Via di Centocelle, 301
00175, Rome

LUXEMBOURG
Voir Belgique

NORVEGE
Norwegian Defence Research
Establishment
Attn: Biblioteket
P.O. Box 25
NO-2007 Kjeller

PAYS-BAS
Royal Netherlands Military
Academy Library
P.O. Box 90.002
4800 PA Breda

POLOGNE
Centralna Biblioteka Wojskowa
ul. Ostrobramska 109
04-041 Warszawa

PORTUGAL
Estado Maior da Força Aérea
SDFA – Centro de Documentação
Alfragide
P-2720 Amadora

REPUBLIQUE TCHEQUE
Vojenský technický ústav s.p.
CZ Distribution Information Centre
Mladoboleslavská 944
PO Box 18
197 06 Praha 9

ROUMANIE
Romanian National Distribution
Centre
Armaments Department
9-11, Drumul Taberei Street
Sector 6
061353 Bucharest

ROYAUME-UNI
Dstl Records Centre
Rm G02, ISAT F, Building 5
Dstl Porton Down
Salisbury SP4 0JQ

SLOVAQUIE
Akadémia ozbrojených síl gen.
M.R. Štefánika, Distribučné a
informačné stredisko STO
Demänová 393
031 01 Liptovský Mikuláš 1

SLOVENIE
Ministry of Defence
Central Registry for EU & NATO
Vojkova 55
1000 Ljubljana

TURQUIE
Milli Savunma Bakanlığı (MSB)
ARGE ve Teknoloji Dairesi
Başkanlığı
06650 Bakanliklar – Ankara

AGENCES DE VENTE

The British Library Document
Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
ROYAUME-UNI

Canada Institute for Scientific and
Technical Information (CISTI)
National Research Council Acquisitions
Montreal Road, Building M-55
Ottawa, Ontario K1A 0S2
CANADA

Les demandes de documents STO, RTO ou AGARD doivent comporter la dénomination « STO », « RTO » ou « AGARD » selon le cas, suivie du numéro de série (par exemple AGARD-AG-315). Des informations analogues, telles que le titre et la date de publication sont souhaitables. Des références bibliographiques complètes ainsi que des résumés des publications STO, RTO et AGARD figurent dans le « NTIS Publications Database » (http://www.ntis.gov).

NORTH ATLANTIC TREATY ORGANIZATION

BP 25
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@cso.nato.int

SCIENCE AND TECHNOLOGY ORGANIZATION

DISTRIBUTION OF UNCLASSIFIED STO PUBLICATIONS

AGARD, RTO & STO publications are sometimes available from the National Distribution Centres listed below. If you wish to receive all STO reports, or just those relating to one or more specific STO Panels, they may be willing to include you (or your Organisation) in their distribution.

STO, RTO and AGARD reports may also be purchased from the Sales Agencies listed below.

Requests for STO, RTO or AGARD documents should include the word 'STO', 'RTO' or 'AGARD', as appropriate, followed by the serial number. Collateral information such as title and publication date is desirable.

If you wish to receive electronic notification of STO reports as they are published, please visit our website (http://www.sto.nato.int/) from where you can register for this service.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Royal High Institute for Defence – KHID/IRSD/RHID
Management of Scientific & Technological Research for Defence, National STO Coordinator
Royal Military Academy – Campus Renaissance
Renaissancelaan 30
1000 Brussels

BULGARIA
Ministry of Defence
Defence Institute "Prof. Tsvetan Lazarov"
"Tsvetan Lazarov" bul no.2
1592 Sofia

CANADA
DSTKIM 2
Defence Research and Development Canada
60 Moodie Drive (7N-1-F20)
Ottawa, Ontario K1A 0K2

CZECH REPUBLIC
Vojenský technický ústav s.p.
CZ Distribution Information Centre
Mladoboleslavská 944
PO Box 18
197 06 Praha 9

DENMARK
Danish Acquisition and Logistics Organization (DALO)
Lautrupbjerg 1-5
2750 Ballerup

ESTONIA
Estonian National Defence College
Centre for Applied Research
Riia str 12
Tartu 51013

FRANCE
O.N.E.R.A. (ISP)
29, Avenue de la Division Leclerc – BP 72
92322 Châtillon Cedex

GERMANY
Streitkräfteamt / Abteilung III
Fachinformationszentrum der Bundeswehr (FIZBw)
Gorch-Fock-Straße 7
D-53229 Bonn

GREECE (Point of Contact)
Defence Industry & Research General Directorate, Research Directorate
Fakinos Base Camp, S.T.G. 1020
Holargos, Athens

HUNGARY
Hungarian Ministry of Defence
Development and Logistics Agency
P.O.B. 25
H-1885 Budapest

ITALY
Ten Col Renato NARO
Capo servizio Gestione della Conoscenza
F. Baracca Military Airport "Comparto A"
Via di Centocelle, 301
00175, Rome

LUXEMBOURG
See Belgium

NETHERLANDS
Royal Netherlands Military Academy Library
P.O. Box 90.002
4800 PA Breda

NORWAY
Norwegian Defence Research Establishment, Attn: Biblioteket
P.O. Box 25
NO-2007 Kjeller

POLAND
Centralna Biblioteka Wojskowa
ul. Ostrobramska 109
04-041 Warszawa

PORTUGAL
Estado Maior da Força Aérea
SDFA – Centro de Documentação
Alfragide
P-2720 Amadora

ROMANIA
Romanian National Distribution Centre
Armaments Department
9-11, Drumul Taberei Street
Sector 6
061353 Bucharest

SLOVAKIA
Akadémia ozbrojených síl gen M.R. Štefánika, Distribučné a informačné stredisko STO
Demänová 393
031 01 Liptovský Mikuláš 1

SLOVENIA
Ministry of Defence
Central Registry for EU & NATO
Vojkova 55
1000 Ljubljana

SPAIN
Área de Cooperación Internacional en I+D
SDGPLATIN (DGAM)
C/ Arturo Soria 289
28033 Madrid

TURKEY
Milli Savunma Bakanlığı (MSB)
ARGE ve Teknoloji Dairesi Başkanlığı
06650 Bakanliklar – Ankara

UNITED KINGDOM
Dstl Records Centre
Rm G02, ISAT F, Building 5
Dstl Porton Down, Salisbury SP4 0JQ

UNITED STATES
Defense Technical Information Center
8725 John J. Kingman Road
Fort Belvoir, VA 22060-6218

SALES AGENCIES

The British Library Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
UNITED KINGDOM

Canada Institute for Scientific and Technical Information (CISTI)
National Research Council Acquisitions
Montreal Road, Building M-55
Ottawa, Ontario K1A 0S2
CANADA

Requests for STO, RTO or AGARD documents should include the word 'STO', 'RTO' or 'AGARD', as appropriate, followed by the serial number (for example AGARD-AG-315). Collateral information such as title and publication date is desirable. Full bibliographical references and abstracts of STO, RTO and AGARD publications are given in "NTIS Publications Database" (http://www.ntis.gov).

ISBN 978-92-837-2253-3