NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

SHOOT THE HORSE AND BUILD A BETTER BARN DOOR: EXPLORING THE POTENTIAL FOR A SUPERFORECASTING METHODOLOGY TO STRENGTHEN THE DHS LEADERSHIP SELECTION PROCESS

by

Ronald Dorman

December 2018

Thesis Advisor: Lauren Wollman (contractor)
Co-Advisor: Douglas J. MacKinnon

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** December 2018	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
SHOOT THE HORSE AND BUILD A BETTER BARN DOOR: EXPLORING THE POTENTIAL FOR A SUPERFORECASTING METHODOLOGY TO STRENGTHEN THE DHS LEADERSHIP SELECTION PROCESS

5. FUNDING NUMBERS

6. AUTHOR(S) Ronald Dorman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)

Over the course of several years, the Department of Homeland Security (DHS) has worked diligently to improve the quality of its leaders. Such efforts have focused almost exclusively on initiating or expanding programs related to leadership development. To date, the impact of that exertion might be charitably described as tepid. While the issues associated with existing leaders have received ample attention, the selection process that precipitated them has not. This gap represents an opportunity to explore a nascent space and suggest new solutions that target the problem at the source. This thesis examines the process of leadership selection at a network level and finds several systemic problems related to measurement, structure, and decision-making. These problems bear a striking resemblance to those observed in the intelligence community and its ability to accurately predict complex future geopolitical events. One method that has dramatically improved the accuracy of geopolitical predictions is superforecasting. At its core, leadership selection is a prediction or a forecast. It is an educated but nonetheless imperfect best guess about how a candidate observed today will perform tomorrow. These features collectively suggest a novel question. Could DHS use a superforecasting methodology to improve its leadership selection process? This thesis follows the progression of that question to an unexpected destination and offers several concrete recommendations.

14. SUBJECT TERMS
superforecasting, leadership, prediction, intuition, accountability, recognition, flow state, closed data loop, heuristics, cognitive bias, judgment, decision making, diversity, crowdsourcing, groupthink, Tversky, Kahneman, Tetlock, homeland security, viewpoint survey, DHS

15. NUMBER OF PAGES
139

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

SHOOT THE HORSE AND BUILD A BETTER BARN DOOR: EXPLORING THE POTENTIAL FOR A SUPERFORECASTING METHODOLOGY TO STRENGTHEN THE DHS LEADERSHIP SELECTION PROCESS

Ronald Dorman
Deportation Officer, U.S. Immigration and Customs Enforcement,
Department of Homeland Security
BA, University of North Carolina at Wilmington, 1995

Submitted in partial fulfillment of the
requirements for the degree of

**MASTER OF ARTS IN SECURITY STUDIES
(HOMELAND SECURITY AND DEFENSE)**

from the

**NAVAL POSTGRADUATE SCHOOL
December 2018**

Approved by: Lauren Wollman
Advisor

Douglas J. MacKinnon
Co-Advisor

Erik J. Dahl
Associate Chair for Instruction
Department of National Security Affairs

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Over the course of several years, the Department of Homeland Security (DHS) has worked diligently to improve the quality of its leaders. Such efforts have focused almost exclusively on initiating or expanding programs related to leadership development. To date, the impact of that exertion might be charitably described as tepid. While the issues associated with existing leaders have received ample attention, the selection process that precipitated them has not. This gap represents an opportunity to explore a nascent space and suggest new solutions that target the problem at the source. This thesis examines the process of leadership selection at a network level and finds several systemic problems related to measurement, structure, and decision-making. These problems bear a striking resemblance to those observed in the intelligence community and its ability to accurately predict complex future geopolitical events. One method that has dramatically improved the accuracy of geopolitical predictions is superforecasting. At its core, leadership selection is a prediction or a forecast. It is an educated but nonetheless imperfect best guess about how a candidate observed today will perform tomorrow. These features collectively suggest a novel question. Could DHS use a superforecasting methodology to improve its leadership selection process? This thesis follows the progression of that question to an unexpected destination and offers several concrete recommendations.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF ACRONYMS AND ABBREVIATIONS

COS	Center for Open Science
CV	curriculum vitae
DHS	Department of Homeland Security
DIA	Defense Intelligence Agency
DNI	Director of National Intelligence
EPJ	expert political judgment
FEVS	Federal Employee Viewpoint Survey
GAO	Government Accountability Office
GCA	general cognitive ability
GJP	Good Judgement Project
HR	human resources
IARPA	Intelligence Advanced Research Projects Activity
IC	intelligence community
JDM	judgment and decision making
JSOC	Joint Special Operations Command
JSU	Jasper State University
KSA	knowledge/skill/achievement
KT	Daniel Kahneman and Amos Tversky
MAT	Miller Analogies Test
OIG	Office of Inspector General
OPM	Office of Personnel Management
RPD	recognition-primed decision-making
SIOP	Society for Industrial and Organizational Psychology
STEM	science, technology, engineering, and mathematics
USAFA	U.S. Air Force Academy
VAM	value added model

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

…Not all those who wander are lost[1]

—Author J.R.R. Tolkien

One of the most enduring criticisms leveled at the Department of Homeland Security (DHS) over the years has been the performance of its leaders. DHS has responded to those criticisms by initiating or expanding programs intended to improve leadership at all levels by focusing on the development of existing leaders. Although these measures have resulted in marginal improvements in some sub-components, DHS remains entrenched in the bottom tier of every government scale of organizational health, a position it has maintained since 2010. While the issues associated with existing leaders have received ample attention, the selection process that precipitated them has not. This gap represents an opportunity to explore a nascent space and suggest new solutions to target the problem at its source.

The DHS leadership selection process relies on candidate storytelling and the ability of experts to intuit the best candidate for a leadership position when they see one. The data to support the belief that experts can intuitively identify leaders is, quite simply, not good. The DHS selection process also fails to measurably correlate predictions with outcomes. In other words, after a promotion decision is made, no one can answer a simple yet vital question: Was that a good decision? Without an objectively accurate answer, officials cannot assess the performance of the selection process to identify errors, make corrections, and produce better subsequent outcomes. Systems that do not objectively compare expected outcomes against actual outcomes are open data loops. They do not produce good outcomes.

In 2011, in response to a string of catastrophic intelligence failures, the U.S. intelligence community (IC) launched a multi-year public forecasting tournament designed to discover better methods for predicting complex future geopolitical events. The winning

[1] J.R.R. Tolkien, *The Fellowship of the Ring* (New York: Ballantine Books, 1977), 231.

team was led by social researcher Philip Tetlock, and the revolutionary method Tetlock created to beat the competition was *superforecasting*.

So, DHS has a leadership problem and the IC has an analysis problem. At its core, leadership selection is really just a prediction, or a forecast. It is an educated but nonetheless imperfect best guess about how a candidate observed today will perform tomorrow. That shared characteristic, as well as the causal overlaps for the fundamental flaws observed in both domains, suggest a novel possibility. Could DHS use a superforecasting methodology to improve its leadership selection predictions in the same manner that Tetlock used it to improve geopolitical predictions?

The answer to that question resides within the thesis. The first chapter contains the problem statement, scope, methodology and assumptions, as well as a comprehensive literature review covering cognitive bias, leadership, and superforecasting. In Chapter II, the literatures are used to disassemble Tetlock's superforecasting process into its constituent parts, and place each piece under a microscope to blueprint foundations, form, and function in detail. What is it, where did it originate, how does it work, why does it work, and what is its role in relation to the other parts in the cycle of operation? The parts are then virtually re-assembled and animated to help the reader visualize the entire process, understand where the data originates, how it is processed and measured, and how feedback is used to refine performance. Chapter III uses the knowledge gained to envision a new Superforecasting process, purpose-built for promotion, and installed in the most advantageous environment imaginable to test the writer's belief that it *might* work, even if only under optimal conditions. Chapter IV assumes a Red Team role and searches for logic gaps or other vulnerabilities in the hypothesis to invalidate the concept. Chapter V recounts the journey to an unexpected conclusion, the insights gained along the way, and opportunities for the next expedition. Ultimately, this thesis identifies systemic flaws in the DHS leadership selection process, and offers several concrete recommendations that can be implemented to produce better organizational outcomes. It also lays a crumb trail for others to follow, and potentially build upon.

Within the confines of a conventional thesis, the purpose of an Executive Summary is to provide the reader with a standalone version that condenses the larger document into

a convenient travel-sized package. This is not a conventional thesis. This is a thought experiment. It is designed to allow the reader to experience the progression of the question through the writer's eyes, stumbles and blunders included. A more complete synopsis would spoil the expedition.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

This is for Katie.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. FRAMING THE QUESTION: ADMIRAL NELSON, GENERAL MCCHRYSTAL, AND THE FRUSTRATING GAP

In *Team of Teams*, General Stanley McChrystal chronicled the encouragement and inspiration he drew from Nelson's improbable defeat of Napoleon's overwhelmingly superior armada.[1] Nelson's gambit consisted of abandoning traditional naval doctrine and its dependency on linear orders radiating from the Admiral to the Captain of each ship of the fleet during battle.[2] Receiving and relaying those orders had historically required ships to fight in a straight line to maintain sight of one another. In lieu of that structure, Nelson empowered his captains to act autonomously, and bring the fight to the enemy as each thought best. This strategy allowed Nelson to drive his ships between Napoleon's traditionally arrayed fleet, thereby interrupting their own command lines and capitalizing on the ensuing chaos. When the acrid plumes were blown away, Britain's bloody fleet stood victorious. The gambit worked.

In analyzing Nelson's success, McChrystal suggests that its central theme is historically misunderstood. Many believe it highlights the virtue of deceit and surprise in combat. The true significance, McChrystal asserts, was Nelson's appreciation of "an organizational culture that rewarded individual initiative and critical thinking, as opposed to simple execution of commands."[3] Nelson's Captains did not magically emerge the day before the battle. They were the product of a clear understanding of the pivotal role they would one day play, and a careful process reflecting that understanding, which began years before the battle.

While leading the Joint Special Operations Command (JSOC) in the fight against Al Qaida, McChrystal found himself looking across the battlefield through Napoleon's

---

[1] Stanley McChrystal et al., *Team of Teams: New Rules of Engagement for a Complex World*, 1st ed. (New York: Portfolio, 2015), 28–31.

[2] Adam Nicolson, *Seize the Fire: Heroism, Duty, and Nelson's Battle of Trafalgar* (New York: Harper Perennial, 2006).

[3] McChrystal et al., *Team of Teams*, 31.

eyes. JSOC had tremendous technological and personnel advantages over the enemy. Yet McChrystal's team had failed to gain traction in the conflict because they were dependent upon a top-down leadership structure that favored predictability over adaptability. Al Qaida was confoundingly unpredictable. McChrystal responded by transforming his organization and re-envisioning the role of leadership to foster the holistic knowledge and trust required to facilitate delegated ground-up decision making in response to a complex environment.

The stories of Nelson and McChrystal contain many similarities. Both identify good leaders as the lynchpins upon which organizational success depends. Both likewise demonstrate the benefits of tailoring leadership structures and roles to achieve an organization's goals. However, both narratives frustratingly fail to ask and answer one intrinsic question. How do you actually go about selecting those leaders in the first place? What would have happened on October 20, 1805, if Nelson had stood upon the dais and asked for volunteers to captain his ships, then facing the crowd of sailors with upraised hands, reached into his pocket and fished out a quarter?

B. PROBLEM STATEMENT

One of the most enduring criticisms leveled at the Department of Homeland Security (DHS) over the years has been the performance of its leaders, and the data supporting those criticisms is compelling. The most frequently cited indication of dysfunction originates from the Federal Employee Viewpoint Survey (FEVS), conducted annually by the Office of Personnel Management (OPM). Since 2002, OPM has administered the survey to both full- and part-time federal employees at every level of government to measure their "perceptions of whether, and to what extent, conditions characteristic of successful organizations are present in their agencies."[4] FEVS is the primary instrument used by the government to measure organizational performance, identify problems, and track outcomes. It is designed using "well-established survey methods that meet the highest professional standards" and produces data with a one-percent

[4] "About the Federal Employee Viewpoint Survey," U.S. Office of Personnel Management, accessed March 9, 2018, https://www.opm.gov/fevs/about.

margin of error.[5] The collective output of the survey over time has reflected a DHS workforce that is disproportionally disengaged when compared with all other federal organizations.

The engagement variable is significant. While many factors can influence how an organization performs, scholars generally believe the most consistent predictor is employee engagement, described by Robinson, Perryman, and Hayday as "a positive attitude held by an employee towards an organization and its value. An engaged employee is aware of business context, and works with colleagues to improve performance within the job for the benefit of the organization."[6]

Over the course of several years, DHS diligently researched the engagement problem identified by FEVS to identify root causes. It commissioned multiple studies, created steering committees and focus groups, conducted internal surveys, and solicited candid feedback from employees in town hall meetings across the country.[7] DHS analyzed the cumulative data from its research and concluded that employee disengagement was the product of poor leadership.[8] That finding is supported by the work of prominent researchers who have likewise held that employee engagement is driven by leadership.[9] An organization's performance rises and falls based on employee engagement, and

---

[5] Doris Hausser, "Understanding the Federal Employee Viewpoint Survey" (working paper, National Academy of Public Administration, 2018), 1, https://www.napawash.org/uploads/AWP_2_Understanding_the_Federal_Employee_Viewpoint_Survey.pdf.

[6] Dilys Robinson, Sarah Perryman, and Sue Hayday, *The Drivers of Employee Engagement*, Report 408 (Brighton, UK: Institute for Employment Studies, 2004), http://www.employment-studies.co.uk/system/files/resources/files/408.pdf.

[7] U.S. Office of the Inspector General, *Major Management and Performance Challenges Facing the Department of Homeland Security* (Washington, DC: U.S. Office of the Inspector General, 2016), 11, https://www.oig.dhs.gov/sites/default/files/assets/2017/OIG-17-08-Nov16.pdf; Office of the Chief Human Capital Officer, *2016 Accomplishments Report* (Washington, DC: Department of Homeland Security, 2016), 28–36; Jerry Markon, "DHS Studies Its Endless Morale Problems, Then Studies Them Some More," *Washington Post*, February 20, 2015, sec. Politics, https://www.washingtonpost.com/politics/homeland-security-has-done-little-for-low-morale-but-study-it--repeatedly/2015/02/20/f626eba8-b15c-11e4-886b-c22184f27c35_story.html.

[8] Office of the Chief Human Capital Officer, *2016 Accomplishments Report*, 31.

[9] Robert Hogan and Robert B. Kaiser, "What We Know about Leadership," *Review of General Psychology* 9, no. 2 (2005): 169–80, https://doi.org/10.1037/1089-2680.9.2.169; James K. Harter et al., *The Relationship between Engagement at Work and Organizational Outcomes* (Washington, DC: Gallup, 2016), http://www.workcompprofessionals.com/advisory/2016L5/august/MetaAnalysis_Q12_ResearchPaper_0416_v5_sz.pdf.

engagement is moderated by leadership.[10] The tool that the federal government relies on to measure organizational performance has consistently identified DHS as a poor performer: DHS has a persistent leadership problem.

That topic has been the subject of reports issued by Congress, the Government Accountability Office (GAO), the Office of Inspector General (OIG), think tanks, and watchdog groups, as well as academic papers, public speeches, and news articles.[11] The external perspectives of the problem are perhaps best encapsulated by a starkly worded OIG memorandum from November 2016. It concluded that DHS leadership challenges were attributable to the repeated failure of leaders to provide meaningful guidance to the workforce, poor communication between leaders and staff, a lack of leader accountability, the absence of effort to build employee relationships, and insufficient attention paid to training and developing the DHS leadership corps.[12]

In response, DHS has initiated or expanded programs intended to improve leadership at all levels by focusing on the development of existing leaders via training and

---

[10] U.S. Merit Systems Protection Board, *Call to Action: Improving First-Level Supervision of Federal Employees* (Washington, DC: U.S. Merit Systems Protection Board, 2010), 4, https://www.mspb.gov/MSPBSEARCH/viewdocs.aspx?docnumber=516534&version=517986&application=ACROBAT.

[11] Tom Colburn, *A Review of the Department of Homeland Security's Missions and Performance* (Washington, DC: Committee on Homeland Security and Governmental Affairs, 2015), http://www.hsgac.senate.gov/download/?id=B92B8382-DBCE-403C-A08A-727F89C2BC9B; Mark T. Kaminsky, "Effective Selection: A Study of First-Line Supervisor Selection Processes in the Department of Homeland Security" (master's thesis, Naval Postgraduate School, 2011), http://www.dtic.mil/docs/citations/ADA543301; Markon, "DHS Studies Its Endless Morale Problems, Then Studies Them Some More"; Jeffrey M. Miller, *Rescuing Tomorrow Today: Fixing Training and Development for DHS Leaders*, Accession Number: AD1029855 (Monterey, CA: Naval Postgraduate School, 2016), http://www.dtic.mil/docs/citations/AD1029855; "Agency Report: Department of Homeland Security," Partnership for Public Service, accessed September 29, 2017, http://bestplacestowork.org/BPTW/rankings/detail/HS00; Jerry Markon, "Homeland Security Ranks Dead Last in Morale—Again—but Jeh Johnson's Morale Is High," *Washington Post*, September 29, 2015, https://www.washingtonpost.com/news/federal-eye/wp/2015/09/29/dhs-disappointed-by-latest-low-morale-scores-vows-to-keep-trying/?utm_term=.c81030c66e3b.

[12] U.S. Office of the Inspector General, *Major Management and Performance Challenges Facing the Department of Homeland Security*, 3.

mentorship.[13] In October 2017, then-acting Secretary Elaine Duke announced a "DHS leadership year" to promote awareness about available development resources, and the critical role leadership plays in mission success.[14] Although these measures have resulted in marginal improvements, DHS remains entrenched in the bottom tier of every government scale of organizational health since 2010.[15]

C. A GAMBLING FALLACY

While the issues associated with existing leaders have received ample attention, the selection process that precipitated them has not.[16] This gap represents an opportunity to explore a nascent space and suggest new solutions to target the problem at its source. Good leaders are critical to DHS, and some candidates are more suitable than others.[17] Were this not true, DHS would forego the time and expense of a formal selection process and simply choose its leaders by rolling dice or drawing straws. The irony is, *that* might be exactly what DHS is unwittingly doing now.

> The most common misunderstanding about science is that scientists seek and find truth. They don't—they make and test models. . . . Building models is very different from proclaiming truths.[18]

---

[13] "DHS Leader Development Program | Office of Leadership (CG-12C)," U.S. Coast Guard, accessed March 11, 2018, http://www.dcms.uscg.mil/Our-Organization/Assistant-Commandant-for-Human-Resources-CG-1/Civilian-Human-Resources-Diversity-and-Leadership-Directorate-CG-12/Office-of-Leadership-CG-12C/DHS-Leader-Development/; Government Accountability Office, *DHS Training: Improved Documentation, Resource Tracking, and Performance Measurement Could Strengthen Efforts*, GAO-14-688 (Washington, DC: Government Accountability Office, 2014), 10, http://www.gao.gov/products/GAO-14-688.

[14] "DHS Leadership Year," Department of Homeland Security, December 6, 2017, https://www.dhs.gov/dhs-leadership-year.

[15] "Unlocking Federal Talent," Office of Personnel Management, accessed March 11, 2018, https://www.unlocktalent.gov/employee-engagement; "Best Places to Work Agency Rankings," Partnership for Public Service, accessed March 11, 2018, http://bestplacestowork.org/BPTW/rankings/overall/large.

[16] National Academy of Public Administration, *Building a 21st Century SES: Ensuring Leadership Excellence in Our Federal Government* (Washington, DC: National Academy of Public Administration, 2017), 163, https://www.napawash.org/uploads/Academy_Studies/Building-a-21st-Century-SES-3.17.2017.pdf.

[17] Shelly Kirkpatrick and Edwin Locke, "Leadership: Do Traits Matter?," *The Executive* 5, no. 2 (May 1991): 48, https://sites.fas.harvard.edu/~soc186/AssignedReadings/Kirkpatrick-Traits.pdf.

[18] Neil Gershenfeld, "Truth is a Model," in *This Will Make You Smarter: New Scientific Concepts to Improve Your Thinking*, ed. John Brockman (New York: Harper Perennial, 2012), 72–73.

While varying somewhat between sub-components, the DHS leadership selection process relies essentially on self-representations made by candidates in the form of resumes, responses to knowledge/skill/achievement (KSA) questionnaires, and oral interviews.[19] In effect, selection decisions are based on storytelling, or how convincingly candidates can portray ideal versions of themselves to experts, and the ability of experts to intuit the best candidate for a leadership position when they see one.

The flawed belief that experts can intuitively identify leaders is not new. Nobel Laureate and psychologist Daniel Kahneman describes the same fallacies at work in the Israeli Army's overconfident reliance on intuition to select officer candidates. Up until the mid-1950s, Israel selected officers based solely on the opinions of expert evaluators following brief interviews and observations. Kahneman tested the accuracy of those predictions and found that they were about as reliable as random chance.[20] A coin toss would have been faster, cheaper, and equally effective. That finding revolutionized the Israeli Army and kicked off a lifetime of research dedicated to studying cognitive bias and flawed heuristics in decision making.[21]

The validity of expert intuition was also the subject of a groundbreaking research project by noted psychologist and political scientist Philip Tetlock. In 1984, Tetlock began a sweeping 20-year study to quantitatively test the predictive accuracy of well-credentialed experts. When the results were tallied, the experts were found to be no more accurate than a dart-throwing monkey.[22] While Tetlock's study relates to expert geopolitical predictions rather than expert leadership predictions, the literature suggests the net result is the same. Expert intuition is not a reliable basis for decisions in domains of complexity.

---

[19] Kaminsky, "Effective Selection," 36–38.

[20] Daniel Kahneman, "Don't Blink! The Hazards of Confidence," *New York Times*, sec. Magazine, October 19, 2011, https://www.nytimes.com/2011/10/23/magazine/dont-blink-the-hazards-of-confidence.html.

[21] Daniel Kahneman and Amos Tversky, "Intuitive Prediction: Biases and Corrective Procedures," June 1977, http://www.dtic.mil/docs/citations/ADA047747; Amos Tversky and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science* 185, no. 4157 (1974): 1124–31, http://www.jstor.org/stable/1738360.

[22] Philip E. Tetlock, *Expert Political Judgment: How Good Is It? How Can We Know?*, 1st. pbk. ed. (Princeton, NJ: Princeton University Press, 2006), 20.

The DHS leadership selection process also fails to measurably correlate predictions (how an expert thought a candidate would perform as a leader if promoted) with outcomes (how that leader actually performed after promotion). This means that after a promotion decision is made, no one can answer a simple yet vital question: Was that a good decision? Without an objectively accurate answer, officials cannot assess the performance of the selection process to identify errors, make corrections, and produce better subsequent outcomes.

The problem associated with measuring the performance of a leader is likewise not new, and relates to another deceptively simple question: How do you define leadership? Decades of research has thus far failed to produce a dispositive answer to that question. While the term *leadership* is conversationally ubiquitous, it is not "a scientific term with a formal, standardized definition."[23] The failure to define it has led to an associated inability to measure it. Today, it is impossible for DHS (or anyone else) to measure the performance of a leader in a way that empirically correlates back to the prediction that resulted in a leader's selection.

Even if it were possible to objectively measure a leader, the current DHS performance management system would be insufficient to the task because it finds that 99% of all employees are fully successful in their performance, which makes any search for meaningful distinctions a futile endeavor.[24] It is impossible to succeed if it is impossible to fail. Systems that do not objectively compare expected outcomes against actual outcomes are open data loops.[25] Much like an assembly line that terminates in a dark room, open data loop systems repeat the same process over and over, with no means of identifying flaws to improve performance. DHS uses an open data loop system to select its leaders. They do not produce good outcomes.

---

[23] Victor H. Vroom and Arthur G. Jago, "The Role of the Situation in Leadership," *American Psychologist* 62, no. 1 (2007): 17–24, https://doi.org/10.1037/0003-066X.62.1.17.

[24] Robert Goldenkoff, *Federal Workforce: Sustained Attention to Human Capital Leading Practices Can Help Improve Agency Performance*, GAO-17-627T (Washington, DC: Government Accountability Office, 2017), 13.

[25] Chris Anderson, "Closing the Loop: A Conversation with Chris Anderson," Edge, accessed October 22, 2017, https://www.edge.org/conversation/chris_anderson-closing-the-loop.

> That which is measured improves. That which is measured and reported improves exponentially.[26]

Superforecasting is the term Tetlock coined to describe a revolutionary method for improving intelligence forecasts of complex future events.[27] Superforecasting mitigates the impact of unconscious bias and flawed heuristics by harnessing the aggregated judgments of multiple crowdsourced forecasters in place of individual experts. Forecasters receive training in such areas as bias mitigation, heuristics, groupthink avoidance, and probability. This improves accuracy. Forecasters are placed into loosely connected groups to facilitate collaboration, red teaming, and post-mortem analysis, which improves accuracy. The accuracy of every forecaster's prediction is scored, and the value of a future prediction is weighted (artificially adjusted) based on the forecaster's past performance. This improves accuracy. Scores are used to identify and correct errors in a forecaster's judgment, which also improves accuracy. Finally, scores are published to spur competition for recognition, thereby driving forecasters to work harder, which improves accuracy.

The core question asked by the status quo DHS selection process is "Which of these candidates will be the best leader?" It is an intuitive, subjective prediction made by an expert, or a small group of experts. This thesis explores what would happen if that question were fundamentally changed. What if good leadership were defined not as the presence of isolated characteristics, self-reported and desirable *sui generis*, but rather as the sum total of leader traits, characteristics, or behaviors *that cause a group to perform the way an organization wants it to*? If group performance (which can be measured) became a proxy for leadership performance (which cannot be measured), then the core question asked by the selection process becomes "Which of these candidates will produce the best performing group?" Using a superforecasting methodology to ask that question mitigates the impact of bias inherent in expert intuition. Measuring the difference between predicted outcomes before a promotion and actual outcomes thereafter would produce analyzable data to

---

[26] Mark Joyner, "Pearson's Law and How the New "Trackers" Feature Improve Things "Exponentially," *Simpleology* (blog), November 16, 2011, http://www.simpleology.com/blog/2011/11/pearsons-law-and-how-the-new-trackers-feature-will-improve-things-exponentially.html.

[27] Philip E. Tetlock and Dan Gardner, *Superforecasting: The Art and Science of Prediction* (New York: Random House, 2015).

identify flaws, update beliefs, and improve subsequent predictions. Different organizations need their groups to perform in different ways. A leadership position may require a range of different characteristics depending on how an organization wants a specific group to perform. Likewise, the optimal tools to identify those traits are subject to change. In this way, a superforecasted leadership selection process is modular and can be individually tailored to keep pace with the changing needs of different DHS subcomponents.

While the independent literatures for superforecasting, cognitive bias, decision making, and leadership are robust, the area where those domains potentially overlap is uncharted territory. This thesis seeks to investigate that space and its potential to strengthen DHS by examining leadership selection at the system or network level. At its core, leadership selection is simply a prediction, or a forecast. It is an educated but nonetheless imperfect best guess about how a candidate observed today will perform tomorrow. It is a process predicated on loosely held beliefs that must constantly be tested, analyzed, questioned, measured, and updated with new data.

D. RESEARCH QUESTION

Could DHS use a superforecasting methodology to improve its leadership selection process?

E. LITERATURE REVIEW

1. Cognitive Bias as It Relates to Judgment and Decision-Making

Superforecasting is grounded in multiple interrelated fields of study, including judgment and decision making (JDM), heuristics, accountability theory, risk management, and network theory. Tetlock's first book, *Expert Political Judgment* (EPJ), demonstrated the pervasive and enduring appeal of raw intuition as a flawed basis for expert judgment.[28] That finding served as a springboard to launch a new understanding of how to reliably improve the predictive accuracy of forecasts related to complex future geopolitical

[28] Tetlock, *Expert Political Judgment*.

events.[29] Tetlock's view that poor judgment is rooted in cognitive bias and flawed heuristics is only useful to the extent that it successfully explains why.[30] For an answer to that question, the most authoritative voices are those of psychologists Daniel Kahneman and Amos Tversky (KT), who began a partnership on the subject in 1969.[31]

In broad-brush terms, the collective output of their research is the finding that humans unconsciously bridge gaps between the known and the unknown by adopting beliefs unbound by empirical evidence or logic.[32] These mental shortcuts, or heuristics, are frequently flawed. They are also generally avoidable. KT's research methodology typically involved asking study participants to make subjective estimates about uncertain future events in which some, but not all, variables were presented, and evaluating the manner in which participants cognitively filled in the gaps to form a judgment.[33] Consistently erroneous leaps in logic were examined and subsequently labeled by KT to (hopefully) diminish their impact on future decisions. For example, KT found that humans tend to misinterpret how important a variable is based on how readily it comes to mind. KT named this phenomenon the availability heuristic.[34] One might use it today to explain why an average American would overestimate the odds of being the victim of a natural disaster.[35] While some might hasten to presume that poor judgment correlates to the intellectual sophistication of the forecaster, KT's research demonstrates that intelligent experts are not immune, and in matters involving confidence, are likely even more susceptible to error.[36]

---

[29] Tetlock and Gardner, *Superforecasting*.

[30] Tetlock and Gardner, 295.

[31] Daniel Kahneman, *Thinking, Fast and Slow* (New York: Farrar, Straus and Giroux, 2013), loc. 42 of 9411, Kindle.

[32] Tversky and Kahneman, "Judgment under Uncertainty."

[33] Daniel Kahneman and Amos Tversky, "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology* 3 (1972): 430–54, http://datacolada.org/wp-content/uploads/2014/08/Kahneman-Tversky-1972.pdf.

[34] Daniel Kahneman and Amos Tversky, "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology* 5 (1973): 207–32, https://msu.edu/~ema/803/Ch11-JDM/2/TverskyKahneman73.pdf.

[35] Eleanor Cummins, "There Was another Earthquake in Mexico. Is the World Ending?," *Slate*, September 19, 2017, http://www.slate.com/articles/health_and_science/science/2017/09/this_summer_has_been_an_unending_series_of_natural_disasters.html.

[36] Kahneman and Tversky, "Intuitive Prediction."

Following Tversky's death in 1996, Kahneman continued his work in the field, and in 2013, published a book that parsed the dichotomous relationship between the intuitive and deliberate processes believed to characterize human thought: System 1 and System 2.[37] System 1 is quick, effortless, and responsible for the bulk of human judgments.[38] System 2 is much slower, objectively analytical, and requires effort to initiate.[39] The literature associated with bias and heuristics, both original and derivative, is robust, and the findings have been replicated across multiple disciplines.[40]

This is not to say that it is an uncontested field of research. Perhaps the most vocal critic over the years has been psychologist and social researcher Gerd Gigerenzer, who believed that the sheer number of heuristic flaws identified by KT rendered the collective findings worthless.[41] At one point, Gigerenzer compared KT's work with the interpretative variability of a Rorschach test, which is not an accusation completely devoid of merit.[42] The first verbal volley fired by Gigerenzer sparked an epic nerd war between the two collegiate camps that would find no equal until the bloody Kirk/Picard conflict of 2003.[43]

For a more thoughtful criticism of KT's work, one might turn to Gary Klein, a former research psychologist for the U.S. Air Force. While KT's work focused on the virtue of System 2 thinking and the means to developing its use, Klein took a polar opposite approach. His research engaged the instances in which experts have reliably used intuitive

---

[37] Kahneman, *Thinking, Fast and Slow*, loc. 207.

[38] Kahneman, loc. 268.

[39] Kahneman, loc. 268.

[40] Martie G. Haselton, Daniel Nettle, and Paul W. Andrews, "The Evolution of Cognitive Bias," in *The Handbook of Evolutionary Psychology*, ed. David M. Buss (Hoboken, NJ John Wiley & Sons, 2005), 724–46.

[41] Gerd Gigerenzer, "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases,'" *European Review of Social Psychology*, 2, no. 1 (1991): 83–115.

[42] Gerd Gigerenzer, "On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky," *Psychological Review* 103, no. 3 (1996): 592–96; "The Complete List of Cognitive Biases," Mind Mastery, accessed November 4, 2017, http://www.mind-mastery.com/article/322/The-Complete-List-of-Cognitive-Biases.

[43] Daniel Kahneman and Amos Tversky, "On the Reality of Cognitive Illusions," *Psychological Review* 103, no. 3 (1996): 582–91; El Santo, "Kirk vs. Picard: Who's the Best Starfleet Captain?," *Rooktopia* (blog), May 16, 2013, https://rooktopia.wordpress.com/2013/05/16/kirk-vs-picard-whos-the-best-starfleet-captain/.

judgment to derive accurate solutions at blazing speeds. Chess masters sort through tens of thousands of potential moves to discern an optimal solution in seconds, medical professionals instantly diagnose complex illnesses from a handful of observed symptoms, and firefighters demonstrate an uncanny capacity to recognize danger in the fleeting moments before disaster strikes.[44] Klein dubbed this phenomenon *naturalistic decision making*, and his theory of it suggests that experts make sound judgments based on experience with past outcomes as opposed to an analysis of future alternatives.[45] The genesis of this theory began with a research methodology involving interviews with fire commanders following critical incidents.[46] The consensus among Klein's research subjects was that they did not engage in a deliberate, analytical process, or option-weighing exercise to achieve an acceptable conclusion; instead, they relied on the instant recollection and aggregation of prior experiences and outcomes to guide their actions.[47] Based on his research, Klein believed that training should focus on the repetitive drilling of pattern-matching and recognition exercises, to enable faster intuitive responses.[48]

In 2006, a recommendation was made to the U.S. Army to adopt Klein's recognition-primed decision-making (RPD) conclusions into a new decision model to replace the ponderous seven-step process in place at the time.[49] The National Fire Academy and the U.S. Marine Corps have incorporated RPD-based drills into their respective

---

[44] Adriaan D. de Groot, *Thought and Choice in Chess* (Berlin: Walter de Gruyter GmbH & Co KG, 1978); Beth Crandall and Karen Getchell-Reiter, "Critical Decision Method: A Technique for Eliciting Concrete Assessment Indicators From the Intuition of NICU Nurses," *ANS. Advances in Nursing Science* 16, no. 1 (September 1993): 42–51; Gary Klein, Roberta Calderwood, and Anne Clinton-Cirocco, "Rapid Decision Making on the Fire Ground: The Original Study Plus a Postscript," *Journal of Cognitive Engineering and Decision Making* 4, no. 3 (September 1, 2010): 186–209, https://doi.org/10.1518/155534310X12844000801203.

[45] Gary Klein, "Naturalistic Decision Making," *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, no. 3 (June 2008): 456–60, https://doi.org/10.1518/001872008X288385.

[46] Gary A. Klein, *A Recognition-Primed Decision (RPD) Model of Rapid Decision Making* (New York: Ablex Publishing Corporation, 1993), 138–47.

[47] Klein, 138–47.

[48] Gary A. Klein, *Sources of Power: How People Make Decisions*, 20th ed. (Cambridge, MA: MIT Press, 2017), 42.

[49] Michael Forsyth and David A. Bushey, "The Recognition-Primed Decision Model: An Alternative to the MDMP for GWOT," *Field Artillery*, January 1, 2006, 13.

regimens.[50] Much like KT, however, Klein was not immune to criticism. In a 2004 article for *Military Review*, Dr. Karol Ross et al. asked a series of reasonable questions.[51] What if the experience upon which a future decision is based was not very good, or if the problem set is entirely novel? Those are valid concerns in a complex world. Another good question might be: How credible are the recollections on which RPD research is based, given that they were all made with the benefit of hindsight?[52]

Thus far, the competing scholarship of Kahneman and Klein appears to occupy seats at opposite ends of judgment's ideological table, with intuition and objective deliberation at odds, and mutually exclusive. A closer inspection, however, suggests surprising areas of consensus. Perhaps demonstrating an intellectual maturity that comes with 14 years of diligent study following the Gigerenzer slap fight, Kahneman and Klein took advantage of a tragic incident to explore the domains of their respective approaches to judgment and collaboratively authored a paper, which found surprising areas of consensus.[53]

In 1988, the USS *Vincennes* erroneously launched an Aegis cruise missile at an overhead Iranian commercial airliner.[54] The fallout from that event precipitated a meeting sponsored by the U.S. Navy, which was attended by 30 accomplished researchers in the field of decision making that inaugurated a seven-year study of tactical judgment.[55] The conference put Kahneman and Klein in the same room and afforded them an opportunity to cooperatively explore a salient question: When is expert intuition a reliable basis for

---

[50] Klein, *Sources of Power*, 44.

[51] Karol G. Ross et al., "The Recognition-Primed Decision Model," *Military Review*, August 2004, 6–10.

[52] Neal J. Roese and Kathleen D. Vohs, "Hindsight Bias," *Perspectives on Psychological Science* 7, no. 5 (September 1, 2012): 411–26, https://doi.org/10.1177/1745691612454303.

[53] Daniel Kahneman and Gary Klein, "Conditions for Intuitive Expertise: A Failure to Disagree," *American Psychologist* 64, no. 6 (2009): 515–26, https://doi.org/10.1037/a0016755.

[54] Jeremy R. Hammond, "The 'Forgotten' U.S. Shootdown of Iranian Airliner Flight 655," *Foreign Policy Journal*, July 3, 2017, https://www.foreignpolicyjournal.com/2017/07/03/the-forgotten-us-shoot down-of-iranian-airliner-flight-655%c2%ad/.

[55] Janis A. Cannon-Bowers and Eduardo Salas, eds., *Making Decision Under Stress: Implications for Individual and Team Training*, 1st ed. (Washington, DC: American Psychological Association, 1998).

decision making?[56] They came to a simple answer: It depends on the environment in which those judgments are made.[57]

Experts thrive in high-validity environments where causal cues are readily available and reliably correlate to specific outcomes. Thus, over time, firefighters develop valuable expertise by repetitively experiencing cause and effect events, and therefore, accurately predict outcomes with much greater speed than that required by System 2 analysis.[58] Such knowledge is limited, however, to that specific paradigm. Hence, a firefighter's ability to judge when a burning structure will implode does not transfer to an understanding of when an overstressed bridge will collapse. Conversely, expert intuition performs poorly in low-validity environments, such as those presented by innumerable geopolitical events or financial market outcomes.[59] In such settings, superior results can be obtained by employing a System 2 approach in concert with algorithms to mitigate the impact of bias. Indeed, a broad meta-analytic study cited by both Kahneman and Klein found that algorithms measurably outperform human judgment in complex environments.[60] In the minority of instances in which algorithms did not outperform expert judgment, the results indicate no measurable difference in accuracy outcomes such that, on balance, mechanical tools still produce better outcomes; not perfect, just better.[61] Both camps likewise agree on the impact of luck in decision making and its likelihood of leading an expert to conclude, erroneously, that a good outcome was the product of a good decision.[62] For example, in poker, an expert might win a hand despite an unpredicted turn of the cards. Pre- and post-mortem decision analysis is essential for both methodologies.[63] One curious conclusion of

---

[56] Kahneman and Klein, "Conditions for Intuitive Expertise," 524.

[57] Kahneman and Klein, 523.

[58] Kahneman and Klein, 524.

[59] Kahneman and Klein, 523.

[60] William M. Grove et al., "Clinical Versus Mechanical Prediction: A Meta-Analysis," *Psychological Assessment* 12, no. 1 (2000): 19–30, https://doi.org/10.1037//1040-3590.12.1.19.

[61] Kahneman and Klein, "Conditions for Intuitive Expertise," 525.

[62] Kahneman and Klein, 524–25.

[63] Deborah J. Mitchell, J. Edward Russo, and Nancy Pennington, "Back to the Future: Temporal Perspective in the Explanation of Events," *Journal of Behavioral Decision Making* 2, no. 1 (January 1, 1989): 25–38, https://doi.org/10.1002/bdm.3960020103.

note is the shared belief that medicine is a high-validity environment, and therefore, well-suited for the application of expert judgment.[64] There is a significant body of research pointing to the negative impact of raw intuition on that field and the accuracy improvements realized by applying System 2 methodologies.[65]

The contribution of KT's research to the field of superforecasting is profound. Its collective output defines the flaws in intuitive decision making and charts a cautiously optimistic path toward mitigated effect and reliably improved accuracy; not perfect, just better. This writer can find little fault in Tetlock's final assertion that "the heuristics-and-biases perspective still provides the best first-order approximation of the errors that real-world forecasters make and the most useful guidance on how to help forecasters bring their error rates down."[66]

2. An Academic Perspective of Leadership

> The Seer-Sucker Theory: No matter how much evidence exists that seers do not exist, suckers will pay for the existence of seers.[67]
>
> —Author J. Scott Armstrong

What is leadership? More than a century of rigorous research by psychologists and social scientists has failed to produce a universally accepted definition.[68] In a 2007 NPS thesis, Nola Joyce counted more than 135,000 different definitions of leadership in academic literature.[69] This taxonomic imprecision caused one prominent researcher to

---

[64] Kahneman and Klein, "Conditions for Intuitive Expertise," 524.

[65] Tetlock and Gardner, *Superforecasting*, 25–30.

[66] Tetlock and Gardner, 295.

[67] J. Scott Armstrong, "The Seer-Sucker Theory: The Value of Experts in Forecasting," *Technology Review* 82, no. 7 (June 1980): 16, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=648763.

[68] Mike Stutzman and Tracy K. Tunwall, "Leadership Success or Failure: Understanding the Link between Promotion Criteria and Leader Effectiveness," *Journal of Business and Economics* 4, no. 8 (August 2003): 691, http://173.83.167.93/UploadFile/Picture/2014-6/201461493432288.pdf.

[69] Nola Joyce, "Can You Lead Me Now? Leading in the Complex World of Homeland Security" (master's thesis, Naval Postgraduate School, 2007), 19, https://calhoun.nps.edu/bitstream/handle/10945/3286/07Sep_Joyce.pdf?sequence=1&isAllowed=y.

lament, "there are almost as many definitions of leadership as there are persons who have attempted to define the concept."[70]

In lieu of a comprehensive definition, scholars through the years have filled the void by parsing leadership into narrow subcategories for study. Early research in the mid-1800s began with a trait paradigm focused on heritable qualities.[71] Subsequent work proposed leadership as a function of demographics, skills, and abilities.[72] Looking beyond those factors led to an examination of the topic from a psychological perspective and focused on personality.[73] Other studies concluded that a leader's psychological makeup is less important than the behaviors that she or he exhibits.[74] Still others have seen leadership as situationally dependent, and have instead focused on environmental factors.[75] This integrative void has created a muddy academic landscape in which leadership has been alternately defined or described by credible authorities using a broad range of terms, including influential, transactional, transformational, sensemaking, heroic, charismatic, narcissistic, and structural.[76] Despite the proliferation and lack of cohesion observed in existing constructs, leadership scholars have continued to "create new theories of leadership without attempting to compare and contrast the validity of existing theories."[77] To date, leadership is a term of art, not science.

---

[70] Bernard M. Bass and Ruth Bass, *The Bass Handbook of Leadership: Theory, Research, and Managerial Applications*, 4th ed. (New York: Free Press, 2008), 11.

[71] Francis Galton, *Hereditary Genius: An Inquiry into Its Laws and Consequences*, 2nd ed. (London: Macmillan, 1892), http://galton.org/books/hereditary-genius/text/pdf/galton-1869-genius-v3.pdf.

[72] D. Scott Derue et al., "Trait and Behavioral Theories of Leadership: An Integration and Meta-Analytic Test of Their Relative Validity," *Personnel Psychology* 64, no. 1 (2011): 7–8, http://online library.wiley.com/doi/10.1111/j.1744-6570.2010.01201.x/full; Alice H. Eagly, Mona G. Makhijani, and Bruce G. Klonsky, "Gender and the Evaluation of Leaders: A Meta-Analysis," *Psychological Bulletin* 111, no. 1 (1992): 3–4, https://doi.org/10.1037/0033-2909.111.1.3.

[73] Timothy A. Judge et al., "Personality and Leadership: A Qualitative and Quantitative Review," *Journal of Applied Psychology* 87, no. 4 (2002): 765–80, https://doi.org/10.1037//0021-9010.87.4.765.

[74] Timothy A. Judge and Ronald F. Piccolo, "Transformational and Transactional Leadership: A Meta-Analytic Test of Their Relative Validity," *Journal of Applied Psychology* 89, no. 5 (2004): 755–68, https:// doi.org/10.1037/0021-9010.89.5.755.

[75] Vroom and Jago, "The Role of the Situation in Leadership," 17–24.

[76] Stefan Schulz-Hardt and Felix C. Brodbeck, "Group Performance and Leadership," *An Introduction to Social Psychology*, 2012, 29–41.

[77] Derue et al., "Trait and Behavioral Theories of Leadership," 8.

The failure to define a thing has understandably led to an associated failure to measure it; indeed, measures of leadership effectiveness throughout the literature demonstrate broad interpretative variability.[78] Some researchers have measured leadership effectiveness based on the extent to which someone is perceived to be "leader-like."[79] Other studies have been devoted to identifying the characteristics of successful leaders based on the speed and trajectory of their career progression, on the premise that leaders with successful career tracks are more effective than their slower contemporaries.[80] The correlation between the characteristics that produce successful careers and those that produce successful groups is actually poor.[81] Kahneman might refer to this an example of the *attribute substitution* heuristic, or the unconscious use of a simple problem to solve a complex one.[82]

Does leadership even matter? While still falling short of total agreement, the research does reflect greater consensus across multiple disciplines. Scholars are quick to point out that organizations are complex and subject to "nonlinear interactions among multiple variables in a dynamic system open to outside influences," which is to say that performance is moderated by forces beyond leadership.[83] However, multiple studies have

[78] Derue et al., 9.

[79] Stutzman and Tunwall, "Leadership Success or Failure," 691.

[80] Yair Berson, Orrie Dan, and Francis J. Yammarino, "Attachment Style and Individual Differences in Leadership Perceptions and Emergence," *The Journal of Social Psychology* 146, no. 2 (April 2006): 165–82, https://doi.org/10.3200/SOCP.146.2.165-182.

[81] Stutzman and Tunwall, "Leadership Success or Failure," 691.

[82] Daniel Kahneman and Shane Frederick, "Representativeness Revisited: Attribute Substitution in Intuitive Judgment," in *Heuristics and Biases*, eds. Thomas Gilovich, Dale Griffin, and Daniel Kahneman (Cambridge: Cambridge University Press, 2002), 49–81, https://doi.org/10.1017/CBO9780511808098.004.

[83] Fred Luthans, "Successful vs. Effective Real Managers," *The Academy of Management Executive* 2, no. 2 (1987): 127–32, http://www.jstor.org/stable/4164814; Stutzman and Tunwall, "Leadership Success or Failure," 692.

found a positive relationship between who is in charge of a group and the quality of the group's performance.[84]

The criticality of the role a leader plays within a group is likewise supported from a *Social Identity* perspective.[85] According to Dr. David Brannan:

> The relationship between effective patron leadership and group effectiveness is always a part of in-group cohesion, shared identity, commitment to shared values and goals (Limited Good) and the health of the group positively contributing to the self worth and identity of the individual. This is all in a constant feed back loop that then contributes to more effective and positive perceptions of the group by other individuals. Leadership matters. It matters a lot from my perspective.[86]

Another testament to the efficacy of leadership may be the thriving industry that in recent years has developed around it. In the United States alone, it is a multi-billion-dollar enterprise populated by authors, consultants, motivational speakers, coaches, and various others to help public and private organizations find, evaluate, select, develop, and retain high-performing leaders.[87] Evidence that the leadership industry has produced better leaders or higher-performing organizations as a result of that investment is not good, yet the industry continues to grow.[88] How is it possible that competitive businesses that are

---

[84] Schulz-Hardt and Brodbeck, "Group Performance and Leadership," 49–50; Ralph White and Ronald Lippitt, "Leader Behavior and Member Reaction in Three 'Social Climates,'" in *Group Dynamics: Research and Theory*, ed. Dorwin Philip Cartwright and Alvin Frederick Zander, 3rd ed. (New York: Harper & Row, 1976), 318–35, https://is.muni.cz/el/1451/podzim2013/np2270/um/cartwright_leader0001.pdf; Durga Devi Pradeep and N. R. V. Prabhu, "The Relationship between Effective Leadership and Employee Performance," in *International Conference on Advancements in Information Technology with Workshop of ICBMG IPCSIT Vol. 20 IACSIT Press, Singapore, 198*, vol. 207, 2011, 205–6; Hogan and Kaiser, "What We Know About Leadership," 174–75.

[85] Michael A. Hogg, "A Social Identity Theory of Leadership," *Personality and Social Psychology Review* 5, no. 3 (August 1, 2001): 184–200, https://doi.org/10.1207/S15327957PSPR0503_1.

[86] David Brannan, email message to author, February 8, 2018.

[87] Daniel Howden, "The Illusion of the Leadership Industry," Recruiting Resources: How to Recruit and Hire Better, March 31, 2016, https://resources.workable.com/blog/failures-of-leadership-industry.

[88] "U.S. Employee Engagement 2011–2017," Gallup, July 30, 2017, http://news.gallup.com/poll/214961/gallup-employee-engagement.aspx.

intensely focused on investment returns would continue to fuel an industry that fails to produce positive results?[89] Perhaps it is because no one has been keeping score.

3. Superforecasting

> If I had only followed CNBC's advice, I'd have a million dollars today. Provided I started with 100 million.[90]
>
> —Comedian Jon Stewart

In 1984, Tetlock began a 20-year research project to quantitatively test the predictive accuracy of the kinds of well-credentialed authorities typically called on for advice by government leaders, media outlets, and think tanks. The methodology, although Herculean in task, was simple in design. Tetlock recruited a diverse group of 284 leading experts who derived income from "commenting or offering advice on political and economic trends of significance to the well-being of particular states, regional clusters of states, or the international system as a whole."[91] Those experts were tasked with providing numerical probability estimates of whether very specific future geopolitical or economic events would transpire.[92] In all, more than 80,000 predictions were collected and recorded. Researchers then waited for tomorrow to become yesterday and tallied the scores. The results demonstrated that the public would be better served by consulting a dart-throwing chimp, which is exactly what Tetlock wrote when he published the study in 2005.[93] Today, Tetlock has mixed emotions about the decision to invoke a monkey metaphor because it

---

[89] Jeffrey Pfeffer, "Leadership BS: Fixing Workplaces and Careers One Truth at a Time" (PowerPoint presentation, Stanford Social Innovation Review, Stanford, CA, October 2016), http://www.ssirinstitute.org/wp-content/uploads/2016/06/Pfeffer_Nonprofit-Management-Institute-2016.pdf.

[90] *Daily Show*, directed by Jon Stewart, Comedy Central, March 9, 2009.

[91] Philip E. Tetlock, *Expert Political Judgement: How Good is It? How Can We Know?* (Princeton, NJ: Princeton University Press, 2005), 239.

[92] Tetlock, 20.

[93] Tetlock.

caused many to miss the salient discovery in a groundbreaking body of research: *What* experts think is far less important than *how* they think.[94]

Each of the experts had invested a lifetime of rigorous work and study into mastering one specific area of expertise.[95] In doing so, they became what Tetlock alliteratively described as hedgehogs; or those who "know one big thing, toil devotedly within one tradition, and reach for formulaic solutions to ill-defined problems."[96] Tetlock contrasted hedgehogs with foxes, who know many small things. Hedgehogs are intractably single-minded and unwilling to consider dissonant perspectives or possibilities.[97] They are thus blinded to solutions that may exist beyond their discipline. Such blind spots are more pronounced when hedgehogs are presented with data deemed incongruent with the established values and tenants of their consumers.[98] This unwillingness to subvert narrative for accuracy produces a myopic perspective that renders hedgehogs particularly vulnerable to the cognitive biases that plague humanity.[99]

Tetlock's experts were also handicapped by their dependence on status-conferring recognition from governments and popular media outlets, two institutions frequently "less interested in the dispassionate pursuit of truth than they are in the buttressing of their prejudices."[100] Noted jurist and economist Richard Posner has described such ideological pugilists as "advocates specializing in solidarity, not credence."[101] Experts who dare to stray from established group narratives pay a heavy price. Former White House National

---

[94] Share Parrish, "Philip Tetlock on the Art and Science of Prediction," *The Knowledge Project*, Podcast audio, December 8, 2015. (Writer's note: on one hand, the media latched onto the inflammatory sound bite and provided the research a distribution radius that it otherwise could never have achieved. However, they only paid attention to the identity affirming idea that doctors, diplomats, and scientists were no better than primates.)

[95] Tetlock, *Expert Political Judgement*, 240.

[96] Tetlock, 20.

[97] Bryan Caplan, "Have the Experts been Weighed, Measured, and Found Wanting?" *Critical Review*, November 2, 2007.

[98] Tetlock, *Expert Political Judgement*, 231.

[99] Philip Ball, "The Trouble with Scientists," *Nautilus*, May 14, 2015.

[100] Ball, 232.

[101] Richard Posner, *Public Intellectuals: A Study of Decline: A Critical Analysis* (Cambridge: Harvard University Press, 2001).

Security Advisor H. R. McMaster recently provided a cautionary tale to other professional pundits by deigning to publicly frame Islam as anything other than pure evil.[102] Airtime is, after all, a scarce resource.[103] Are there any predictive realms in which hedgehogs excel? Overconfidence in their discipline means that when they bet, they bet big, making them *really* right on rare occasions.[104] Hedgehogs are also necessarily adept and convincing storytellers with a valuable capacity for asking good questions.[105]

The Iraq intelligence failure in 2003 and others preceding it prompted the Intelligence Advanced Research Projects Activity (IARPA) of the Office of the Director of National Intelligence (DNI) to issue a challenge designed to explore the limits of prediction.[106] The government commissioned five university research teams to compete against an IARPA control team in a forecasting tournament to discover which was best and *why*. Each team was tasked with using nothing but publicly available data to provide answers to probability questions typically entrusted to analysts.[107] Tetlock formed a crowdsourced team of more than 2,800 volunteers and labeled them the Good Judgement Project (GJP).[108] Over the course of four years, GJP provided more than one million predictions in response to roughly 500 questions. Throughout the tournament, Tetlock ran a constant series of internal tests to determine which specific factors resulted in improved accuracy for GJP. Combining those factors that were found to be effective led Tetlock to a replicable recipe for producing analytical forecasts that were far superior to those produced by any other competitor during the IARPA tournament.

---

[102] Aaron Klein, "H.R. McMaster-Endorsed Book Calls Jihad Peaceful, Al-Qaida Terrorism 'Resistance,'" Breitbart, August 18, 2017, https://www.breitbart.com/middle-east/2017/08/18/h-r-mcmaster-endorsed-book-calls-jihad-peaceful-al-qaida-terrorism-resistance/.

[103] Henry Tajfel and John Turner, "An Integrative Theory of Intergroup Conflict," in *The Social Psychology of Intergroup Relations*, eds. William Austin and Stephen Worchel (Monterey, CA: Brooks/Cole, 1979), ch. 3, 33–47.

[104] John Brockman, Russell Weinberger, and Nina Stegeman, "Edge Masterclass 2015: A Short Course in Superforecasting, Class I," Edge, August 24, 2015, https://www.edge.org/conversation/philip_tetlock-edge-master-class-2015-a-short-course-in-superforecasting-class-i.

[105] Brockman et al.

[106] Tetlock and Gardner, *Superforecasting*, 17.

[107] Leonard Mlodinow, "Mindware and Superforecasting," *New York Times*, October 15, 2015.

[108] Tetlock and Gardner, *Superforecasting*, 17.

The fundamental flaws observed within the U.S. intelligence community relating to taxonomy, intuition, bias, and measurement bear a striking resemblance to those associated with the status quo leadership selection process. That possibility may hold promising implications for DHS.

F. PURPOSE, SCOPE, AND LIMITATIONS

This thesis is guided by several assumptions:

- Good leaders are critical to DHS and the execution of its mission. Not all employees would make good leaders, so leadership selection is critical to DHS.
- It is impossible to objectively measure the performance of a leader. Leadership is a term that has never been defined in a way that allows for quantitative comparison or analysis.
- It is possible to objectively measure the performance of a group, so if a leader's performance were to be measured based on the performance of the group for which she or he is responsible, then organizations could compare predicted outcomes to actual outcomes. This would transform leaders and their groups into distributed data-producing sensor arrays, producing analyzable feedback.
- Closed data loop systems designed to compare predicted outcomes with actual outcomes perform better than open data loop systems that do not provide output feedback.
- In complex environments, the accuracy of predictions made by a large, loosely networked group of trained forecasters with diverse backgrounds will outperform predictions made by a small homogenous group of experts.

The information relied upon for this thesis originated from both primary and secondary sources. It synthesized literature and research drawn from multiple fields of

study, as well as published reports and survey data from government agencies, think tanks, and private corporations.

The scope of this thesis is limited to exploring the hypothesis that a superforecasting methodology could improve the DHS leadership selection process. It is important to acknowledge that this thesis cannot conclusively demonstrate that a superforecasting methodology *will* produce measurably better outcomes than the status quo. There are no instruments capable of measuring the accuracy of the current selection process, so a direct comparison is impossible. While the literature demonstrates that an improved process should result in observable changes in DHS (better organizational performance and increased levels of employee engagement etc.), the data for those outcomes reside on a timeline that exceeds the practical limitations of this work. There are no published case studies involving superforecasting and leadership selection. Superforecasting is by nature reliant on the aggregation of predictions made by many forecasters. A valid human study of that kind would require hundreds or thousands of subjects and three years of rigorous data collection. Replication on a small scale with a handful of subjects would produce unreliable data because small samples produce extreme results.[109] Constructing a statistical model to test the hypothesis would be fruitless because human predictions cannot be accurately represented by a number on a graph.

Because these limitations make conventional research methodologies either impractical or impossible, this thesis chose to travel an unconventional path in the form of a *Gedankenerfahrung*, or thought experiment.[110] Thought experiments typically employ subjunctive reasoning to conceptually test an otherwise untestable hypothesis by imagining a set of conditions intended to answer the question "what would happen if?" Thus, an

---

[109] Douglas W. Hubbard, *The Failure of Risk Management: Why It's Broken and How to Fix It*, 1st ed. (Hoboken, NJ Wiley, 2009), 100–101; Amos Tversky and Daniel Kahneman, "Belief in the Law of Small Numbers," *Psychological Bulletin* 76, no. 2 (1971): 105–10, http://stats.org.uk/statistical-inference/TverskyKahneman1971.pdf.

[110] Brendan Bernicker, "Gedankenexperiment|Thought Experiments," Penn State University, February 4, 2016, https://sites.psu.edu/bernickerpassionblog/2016/02/04/gedankenexperiment/; Gino Segre, "Gedankenexperiment," *2011: What Scientific Concept Would Improve Everybody's Cognitive Toolkit?* (blog), 2011, https://www.edge.org/response-detail/10157.

otherwise unfalsifiable fancy may still bear fruit. Notable examples include Schrodinger's Cat, Einstein's relativity theory, and the Turing test.

This conceptual conveyance aspires to allow the reader to follow the evolution of the hypothesis through the writer's eyes (stumbles and blunders included) and proceeds in four chapters. Chapter II uses the established literatures to disassemble Tetlock's superforecasting process into its constituent parts, and places each piece under a microscope to blueprint foundations, form, and function in detail. What is it, where did it originate, how does it work, why does it work, and what is its role in relation to the other parts in the cycle of operation? The parts are then virtually assembled and animated to help visualize the entire process, understand where the data originates, how it is processed and measured, and how feedback is used to refine performance. Chapter III uses the knowledge gained to envision a new superforecasting process, purpose-built for promotion, and installed in the most advantageous environment imaginable to test the writer's belief that it *might* work, even if only under optimal conditions. Chapter IV assumes a Red Team role and searches for logic gaps or other vulnerabilities in the hypothesis to invalidate the concept. Chapter V recounts the journey to an unexpected conclusion, the insights gained along the way, and opportunities for the next expedition.

Predictive models of any type are either accurate or useful, but never both. The model imagined herein is certainly no exception. This thesis is a starting point for exploring the stored potential in a superforecasted leadership selection process, not a destination. It illuminates the challenges and limitations associated with the status quo and explores the potential for the process to evolve into one driven by evidence.

II. UNPACKING TETLOCK: THE NUTS AND BOLTS OF SUPERFORECASTING

A. GALTON'S OX

Sir Francis Galton (1822–1911) was a prolific British scientist and mathematician who made substantive contributions to a disparate range of academic disciplines over the course of his lifetime.[111] One recurrent area of research was *eugenics*, a term Galton coined, which denoted a branch of study dedicated to the hereditary superiority of a minority class of human elites, and their exploitation as breeding stock to improve the Homo sapien herd.[112]

Not surprisingly then, it was with this thought that Galton found himself perusing a county livestock fair one brisk autumn month in 1906.[113] A contest was underway in which attendees were invited to examine a corralled ox, bound for slaughter, and place a wager on the ultimate weight of the beast once it was dispatched and dressed for sale.[114] In exchange for a meager sum, anyone could purchase a ticket and inscribe upon it a best guess that was then submitted for the contest. Following the grizzly deed, the contestants who came closest to the correct answer were rewarded with a prize.[115] Galton's (ultimately misguided) eugenic epistemology left him dubious as to the efficacy of the democratic process; specifically, with regard to a commoner's capacity to make optimal voting decisions about complicated governmental issues. The contest, Galton reasoned, was a reasonable proxy for simpleton suffrage because "the average competitor was probably as well fitted for making a just estimate of the dressed weight of the ox, as an average voter is of judging the merits of most political issues on which he votes, and the variety among

---

[111] "Sir Francis Galton F.R.S," Francis Galton, accessed June 3, 2018, http://galton.org/main.html.

[112] Sara Goering, "Eugenics," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta (Stanford, CA: Metaphysics Research Lab, Stanford University, 2014), https://plato.stanford.edu/archives/fall2014/entries/eugenics/.

[113] Francis Galton, *Memories of My Life* (London: Methuen, 1908), 280–81, http://galton.org/books/memories/galton-memories-1up-v2-300dpi.pdf.

[114] James Surowiecki, *The Wisdom of Crowds*, 1st ed. (New York: Anchor Books, 2005), xi–xiv.

[115] Surowiecki.

the voters to judge justly was probably much the same in either case."[116] The modest entry fee would not bar participation by even the lowest strata of society, and the prospect of prizes would spur contestant rigor. To test this belief, Galton collected all the entry tickets at the conclusion of the contest and analyzed their collective output.

After recording each of the nearly 800 guesses, one might reasonably imagine the quiet, self-satisfied grin that passed Sir Galton's face upon noting that nearly all the contestants were not just wrong, but very wrong. Any confirmational elation was fleeting however, because when Galton averaged the responses, he was forced to acknowledge a disconcerting conclusion. The aggregated wisdom of the common people was nearly perfect. The crowd predicted that the final dressed weight of the ox would be 1,197 pounds. The scales read 1,198. Not only was the crowd extremely accurate, it was more accurate than the best guess of any single member.[117] Galton could have easily clung to preconception and discarded the results, but to his credit, he reported every detail in an article published in 1907. In it, Galton publicly updated his beliefs by concluding, "This result is, I think, more credible to the trustworthiness of a democratic judgment than might have been expected."[118]

Galton's findings were not the product of an isolated instance of chance, and the aggregation method he described has since demonstrated tremendous efficacy across a broad range of problem spaces, including financial markets, election outcomes, geopolitical events, technological innovations, sports betting, and social phenomena.[119] In their book *Blind Man's Bluff*, authors Sherry Sontag and Christopher Drew detail how the wisdom of a crowd was successfully used to pinpoint the location of a lost submarine somewhere in the North Atlantic, as well as a sunken hydrogen bomb off the coast of Spain.[120]

---

[116] Francis Galton, "Vox Populi," *Nature* 75 (1907): 450, https://www.nature.com/nature/journal/v75/n1949/pdf/075450a0.pdf?foxtrotcallback=true.

[117] Tetlock and Gardner, *Superforecasting*, xiii.

[118] Galton, "Vox Populi," 451.

[119] Surowiecki, *The Wisdom of Crowds*, 286–91.

[120] Sherry Sontag, Christopher Drew, and Annette Lawrence Drew, *Blind Man's Bluff: The Untold Story of American Submarine Espionage* (New York: PublicAffairs, 2016), 58–60, 145–50.

The key to crowdsourced intelligence, according to Tetlock, lies in "recognizing that useful information is often dispersed widely, with one person possessing a scrap, another holding a more important piece, a third having a few bits, and so on."[121] In the case of the ox, it is reasonable to presume that some of the contestants were experts in the field (butchers, farmers, purveyors, etc.), yet based on the individual predictions submitted, none of the professionals possessed sufficient information to independently produce the most the most accurate estimate. One contestant may have factored local precipitation levels into their guess, while another may have had some insight into animal husbandry. A third may have known the butcher and how much fat he typically cuts away from the lean. As the number of contestants increases, so too does the small scraps of useful information that each contributes. Valid information increasingly coalesces around the correct answer, while invalid information (or wild guesses) above or below the mark cancel each other out when the collective wisdom of the crowd is aggregated and averaged to produce a prediction. While aggregated predictions from large groups are typically more accurate than the best guess of any individual, this is not always the case. Even within extremely complex domains in which outcomes are uncertain, it is certainly possible (either through luck or skill) for an expert to demonstrate a degree of accuracy in a single instance that borders on clairvoyance. However, it is important to remember that such instances are almost never replicated consistently over time. According to Tetlock, "There will be individuals who beat the group in each repetition, but they will tend to be *different* individuals."[122] In the long run, betting on the group will produce better overall outcomes. The question is this: What kind of group produces *optimal* outcomes?

[121] Tetlock and Gardner, *Superforecasting*, 73.

[122] Tetlock and Gardner, 73.

B. FORECASTER TEAMING

> Madness is the exception in individuals, but the rule in groups.[123]
>
> —Philosopher Friedrich Wilhelm Nietzsche

1. Group Collaboration: Obstacles and Opportunities

For decades following the publication of *Vox Populi*, conventional wisdom held that in order for crowdsourced predictions to produce optimal estimates of future events, it was critical that individual forecasters contributing to the effort produce judgments that were free from external influence. In 2004, James Surowiecki wrote one of the most frequently cited books on crowdsourced prediction, which professed "the best way for a group to be smart is for each person in it to think and act as independently as possible."[124] This is to say that, while predictions should emanate from a large, diverse group of forecasters, the individual contributors of the group should not be afforded an opportunity to work collectively, exchange information, discuss, debate, or influence one another prior to submitting their respective estimates. Surowiecki's assertion rested upon what was at the time believed to be a well-established foundation pertaining to the negative impact of conformity and consensus seeking behavior within groups tasked with working collectively to solve a problem.[125] To grasp the source and nature of Surowiecki's concern, one need only bring to mind a solitary word, *committees*. Sir Barnett Cocks, Clerk of the House of Commons, sardonically referred to such collectives as "a cul-de-sac down which ideas are lured, and then quietly strangled."[126]

---

[123] Friedrich Wilhelm Nietzsche, *Beyond Good and Evil: Prelude to a Philosophy of the Future*, trans. Walter Arnold Kaufmann (New York: Vintage Books, 1989), 90.

[124] Surowiecki, *The Wisdom of Crowds*, XX.

[125] Irving L. Janis, *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*, 2nd ed. (Boston: Houghton Mifflin, 1982), 2–13.

[126] Tam Dalyell, "Westminster Scene," *New Scientist* 60, no. 871 (November 8, 1973): 423, https://books.google.com/books?id=i145R0bZXMYC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false.

2. Repurposing Groupthink

Social psychologist Irving Janis coined the term *groupthink* in 1971 after studying American policy fiascos, such as the Bay of Pigs, Pearl Harbor, the Korean War, and the Vietnam War.[127] Janis sought to understand how otherwise incomparably intelligent individuals could reach such poor collaborative conclusions, and categorized his findings into eight symptoms of congregate dysfunction.[128]

- Invulnerability. Members of decision-making groups feel insulated from any individual blame or accountability for poor judgment because of a tacit understanding that responsibility will ultimately be borne by the group. This belief causes groups to make over confident estimations or ignore contrary indications that exist beyond the established in-group narrative.

- Rationale. Members construct baseless rationalizations when faced with negative feedback, which, if otherwise given consideration, might cause the group to re-evaluate its core assumptions or dominant narrative. As an example, Janis cited President Johnson's advisors and their successive wave of decisions to escalate the intensity of bombings in North Vietnam, in spite of the fact that every previous escalation had proven ineffective.

- Morality. Those afflicted with groupthink blindly accept the inherent morality of the in-group and either ignore or fail to consider the ethical implications of the group's decisions.

- Stereotyping. Members of an in-group stereotype those with divergent or contradictory viewpoints, to diminish the validity of out-group perspectives, such that any serious consideration of an alternative view is unwarranted.

---

[127] Irving Janis, "Groupthink," *Psychology Today*, 84, November 1971, http://agcommtheory.pbworks.com/f/GroupThink.pdf.

[128] Janis, 85–88.

- Pressure. Within the group, dominant voices and narratives emerge, and views begin to coalesce around prevalent opinions or shared illusions. Any members who express countervailing ideas or doubt are pressured by the majority to fall in line and exhibit loyalty to the group.

- Self-Censorship. Those who hold views or opinions, which deviate from the consensus, tend to remain quiet and keep their objections to themselves for the sake of cohesion, as well as the avoidance of pressure.

- Unanimity. Once the dominant viewpoint emerges, members share in the illusion that the group is unanimously behind it. Dissenters censor themselves and each member incorrectly presumes that the resulting silence is synonymous with consent.

- Mindguards. Within the group, individuals emerge who actively work to insulate members from information or opinions that may challenge or discredit the dominant viewpoint.

Imagine the potential implications had Galton's group been gathered together prior to the slaughter and tasked with working cooperatively to produce a best guess. Some forecasters might have resorted to cognitive loafing and failed to devote maximum mental effort to the problem on the belief that others within the group would do the work for them.[129] Within such a group, a small core of thought leaders and influencers might reasonably emerge whom, for reasons related to extraversion, persuasive skill, or perhaps superficially attractive credentials, are able to coalesce the group around a dominant viewpoint.[130] Unanimity and consensus are powerful social forces within a group, which can cause individuals to strive for the least objectionable solution rather than the most accurate solution. Committee members implicitly understand that ultimately the group,

[129] Mark Seidenfeld, "Cognitive Loafing, Social Conformity and Judicial Review of Agency Rulemaking," *Cornell Law Review* 87, no. 2 (January 2002): 511–12, https://doi.org/10.2139/ssrn.280251.

[130] Simon Taggar, Rick Hackew, and Sudhir Saha, "Leadership Emergence in Autonomous Work Teams: Antecedents and Outcomes," *Personnel Psychology* 52, no. 4 (December 1999): 899–926, https://doi.org/10.1111/j.1744-6570.1999.tb00184.x.

rather than its individual members, will be held accountable for the outcome, which decreases individual incentive to strive for accurate results (rigor).

Aggregating a large number of independent predictions has demonstrated utility because, when graphed, predictions predicated upon good judgment cluster around the correct solution and gain resolution when averaged.[131] Predictions resulting from poor judgment tend to be equally distributed, and cancel one another out. The effect dramatically decreases the signal to noise ratio, thereby increasing accuracy, a curious phenomenon that has nonetheless been observed "whether the forecasts are judgmental or econometric or extrapolation."[132] An analogous effect may be experienced simply by donning a pair of *Bose* headphones and thumbing the switch.[133] Allowing forecasters to work collaboratively in groups has been historically problematic because groupthink increases the likelihood of errors, and pulls the bulk of those errors in a single direction. When randomly wrong becomes consistently wrong, averaging does not dampen the noise, it amplifies it, producing static that obscures signal fidelity. For a thorough consideration of groupthink and its impact on decision making within the homeland security enterprise, James Ricciuti's 2014 NPS thesis entitled "Groupthink: A Significant Threat to the Homeland Security of the United States," is a wholly worthwhile resource.[134]

While maintaining independence within crowdsourced groups of forecasters may indeed provide an effective bulwark against the deleterious effects of groupthink, doing so limits a group's prescient potential. Diversity is a silver bullet.[135] The virtue of harnessing

---

[131] Robert T. Clemen and Robert L. Winkler, "Combining Economic Forecasts," *Journal of Business & Economic Statistics* 4, no. 1 (January 1986): 39, https://doi.org/10.2307/1391385.

[132] Robert T. Clemen, "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting* 5, no. 4 (1989): 559.

[133] "How Do Noise Cancelling Headphones Work?—James May's Q&A (Ep 10)—Head Squeeze," BBC Earth Lab, accessed December 26, 2017, https://www.youtube.com/watch?v=VTx4JgYsW5s; "Bose Noise Cancelling Headphones," Bose, accessed July 13, 2018, https://www.bose.com/en_us/products/headphones/noise_cancelling_headphones.html.

[134] James E. Ricciuti, "Groupthink: A Significant Threat to the Homeland Security of the United States" (master's thesis, Naval Postgraduate School, 2014), https://calhoun.nps.edu/bitstream/handle/10945/44650/14Dec_Ricciuti_James.pdf?sequence=1&isAllowed=y.

[135] Lu Hong and Scott E. Page, "Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers," *Proceedings of the National Academy of Sciences* 101, no. 46 (November 16, 2004): 16385–89, https://doi.org/10.1073/pnas.0403723101.

the collective power of a large, diverse group of forecasters is the broad range of knowledge, skills, perspectives, and strategies that can be brought to bear against a complex problem.[136] However, segregation means that the only beneficiary of that cognitive cornucopia is the prediction itself. Maintaining independence prohibits members from exploiting fellow forecasters as a means of improving understanding, generating encouragement, testing assumptions, detecting flaws, or exploring new methodologies. From a social identity perspective, the literature suggests that the desire to perform well in front of peers, and thereby increase one's standing within a group, could be a strong motivational factor.[137] For Tetlock, the prospect of teaming represented a risky roll of the dice, with much to be gained or lost.

Perhaps one of the most notable outcomes of the GJP experiment then was support for the notion that accuracy and independence may comprise a false dichotomy. Through a series of controlled experiments examining outcomes resulting from multiple organizational modalities, GJP researchers observed that under the proper conditions, forecasters could preserve their independence and remain focused on achieving the most accurate prediction, while simultaneously exploiting the benefits of team collaboration and knowledge sharing. At the end of the GJP experiment, researchers analyzed the data produced by the control groups and observed that the accuracy of forecasters who were allowed to work together in teams was 23% better than those working independently, and the performance gap progressively increased over multiple years without regression.[138] In fact, teaming produced results that were superior to any other organizational method.[139]

---

[136] Katherine W. Phillips, "How Diversity Makes Us Smarter," *Scientific American*, accessed May 1, 2018, https://doi.org/10.1038/scientificamerican1014-42; David Rock and Heidi Grant, "Why Diverse Teams Are Smarter," *Harvard Business Review*, November 4, 2016, https://hbr.org/2016/11/why-diverse-teams-are-smarter.

[137] Guido Hertel, Norbert L. Kerr, and Lawrence A. Messé, "Motivation Gains in Performance Groups: Paradigmatic and Theoretical Developments on the Köhler Effect," *Journal of Personality and Social Psychology* 79, no. 4 (2000): 580–601, https://doi.org/10.1037//0022-3514.79.4.580.

[138] Tetlock and Gardner, *Superforecasting*, 201.

[139] Barbara Mellers et al., "Improving the Accuracy of Geopolitical Risk Assessments," in *The Future of Risk Management*, eds. Robert Meyer and Erwann Michel-Kerjan (Philadelphia: University of Pennsylvania Press, forthcoming), 7–8, https://sites.hks.harvard.edu/fs/rzeckhau/Geopolitical%20Risks.pdf.

This finding should not be interpreted as an endorsement of committees. Collaboration can turbo cognitive exhaust into forecasting fuel, but in order for teaming to work successfully, forecasters must learn how to extract maximum benefit from their teammates, and the groups must incentivize selfish altruism.[140] This is to say that forecasters diligently work to improve their group's performance with the understanding that doing so improves their own performance. One factor found to ameliorate groupthink's impact upon the collaborative efficacy of teams over the course of the GJP study was accountability.

3. Optimizing Accuracy via Accountability

In general, the literature suggests that cognitive effort exerted towards a task increases in proportion to the level of anticipated accountability.[141] On that basis, GJP researchers suspected that forecasters could attain the benefits of working collaboratively in teams while still maintaining maximum individual effort by holding individual forecasters, rather than the team to which they were assigned, ultimately accountable. The subsequent question became, accountable for what?

The question was rooted in the difference between *process* and *outcome* accountability.[142] In the case of process accountability, people are judged based upon the quality of the method employed to produce a result.[143] Outcome accountability only considers the quality of the end result, without considering how the result was achieved.[144] Literature on the subject lacked consensus. Some researchers found that outcome

---

[140] Thomas L. Friedman, *Thank You for Being Late: An Optimist's Guide to Thriving in the Age of Accelerations* (New York: Farrar, Straus and Giroux, 2016), loc. 802, Kindle, paraphrasing John Donovan "digital exhaust into digital fuel."

[141] Elizabeth Weldon, "Cognitive Loafing: The Effects of Accountability and Shared Responsibility on Cognitive Effort," *Personality and Social Psychology Bulletin* 14, no. 1 (March 1988): 159–71, http://journals.sagepub.com/doi/pdf/10.1177/0146167288141016.

[142] Jennifer Lerner and Philip Tetlock, "Accounting for the Effects of Accountability," *Psychological Bulletin* 125, no. 2 (1999): 258, http://scholar.harvard.edu/files/jenniferlerner/files/lerner_and_tetlock_1999_pb_paper.pdf.

[143] Lerner and Tetlock, 258.

[144] Lerner and Tetlock, 258.

accountability produced higher levels of performance-inhibiting anxiety.[145] Others held that while process accountability did improve performance when applied to simple unambiguous tasks in which cause and effect were understood, the results were not generalizable across the board.[146] Process accountability might work well for employees of fast food chains (lettuce on top of tomato, not tomato on top of lettuce), but be less effective for employees charged with predicting price futures in volatile commodity markets.[147] GJP researchers noted that most of the research at the time stemmed from studies of single-session experiments involving simple tasks, and did not evaluate a subject's ability to evolve or adapt to a variety of complex challenges over time.[148] Tetlock believed that tasking forecasters to produce the most accurate prediction possible, without stipulating how, would incentivize them to take methodological risks by searching for creative new strategies in response to dynamic problem spaces.[149] Whereas process accountability would produce "consistency with standard practices," the alternative would increase adaptive performance "because outcome goals move people to the novel and unfamiliar, they can gain a sense of enthusiasm, curiosity, and urgency-all of which stimulate exploration and learning."[150] It was a testable hypothesis.

To evaluate the various modalities, GJP researchers publicly recruited 1,850 subjects from across the United States to participate in a one-year forecasting tournament, and randomly assigned them to a series of accountability and collaboration conditions. One might think of it as a smaller experiment tucked within the larger GJP experiment. A zero accountability control group receiving no feedback on either process or outcome was created for benchmark comparison. Subjects were randomly assigned to process, outcome,

---

[145] Robert H. Ashton, "Effects of Justification and a Mechanical Aid on Judgment Performance," *Organizational Behavior & Human Decision Processes* 52, no. 2 (July 1992): 292–306.

[146] Bart de Langhe, Stijn M.J. van Osselaer, and Berend Wierenga, "The Effects of Process and Outcome Accountability on Judgment Process and Performance," *Organizational Behavior and Human Decision Processes* 115, no. 2 (July 2011): 238–52, https://doi.org/10.1016/j.obhdp.2011.02.003.

[147] Rodrigo Nieto-Gomez, email message to author, October 4, 2017.

[148] Chang Welton et al., "Accountability and Adaptive Performance under Uncertainty: A Long-Term View," *Judgment and Decision Making* 12, no. 6 (2017): 611.

[149] Welton et al., 611–13.

[150] Welton et al., 612.

and hybrid accountability conditions. Thirty teams, each comprised of 13 forecasters, were created to work collaboratively, and the residual study subjects were tasked with working independently. Forecasters working within teams were enabled and encouraged to share strategies, justifications, estimates, etc. with their teammates. All the forecasters worked within a custom online platform created by researchers. Those working independently used the online platform to submit forecasts, notate the thought process and justifications that contributed to their forecasts, and receive feedback from researchers. For teams, the platform also served as the means of collaboration via forum conversations, such that team members were loosely networked and never met face to face. Team members could also use the platform to subjectively rate the quality of information, ideas, criticisms, or strategies shared by fellow forecasters. The platform provided researchers with complete visibility in order to measure and record variables including forecaster engagement, knowledge transfer, and belief updating. Every team was organized with a flat leaderless hierarchy that allowed roles and relationships to develop organically within each group.

All of the study participants received a ~ 60-minute on-line training module, dubbed *CHAMPS KNOW*, which served as a best practice guide involving psychological principles and strategies to improve probabilistic reasoning and predictive accuracy. Details of the training module are discussed in a subsequent section. The final portion of the training module detailed how the performance of each forecaster would be evaluated. Those assigned to the process condition were told that they would be judged on the quality of their forecasting method, as well as the extent to which it conformed to best practice guidelines, with no consideration of the outcome of that process. Process forecasters assigned to a team were also judged on their level of engagement and collaboration with teammates, in conformance with established best practice guidelines. Outcome forecasters were informed that their solitary objective was to produce the most accurate predictions possible, without consideration of the process used to obtain that result. Outcome forecasters assigned to teams were encouraged to share and collaborate with their teammates, but the method and frequency of collaboration was left to individual discretion. Forecasters assigned to the hybrid assessment condition were told that both process and

outcome would be equally weighted and scored. This applied to hybrid forecasters working individually or within teams.

Over the course of a year, study participants provided probability estimates for 135 questions regarding the likelihood of complex future geopolitical events coming to pass. The topics were diverse and fluid (currency fluctuations, public health, power politics, etc.), such that a forecaster with a strong knowledge base in one particular discipline or region could not hold sway over the course of the study. Participants could use information or research from any open (unclassified) source. Forecasters in all of the accountability conditions received regular feedback from researchers. Process forecasters received monthly scores reflecting the quality of their methodology. Process scores were the product of a matrix established by researchers which rated factors such as the number of forecasts submitted per question (belief updating), the number of hyperlinks included in a given justification, the number of analytic observations in a justification, and the extent to which justifications conform to best practice guidelines. Process forecasters also routinely received examples of high quality justifications based on CHAMPS KNOW principles to help develop their analytic methodology. Outcome forecasters, conversely, only received scores reflecting the accuracy of their final prediction for each question. Members of the hybrid condition received scores reflecting an equal weight average of outcome and process. All of the participants (excluding the control group) received feedback on the ultimate outcomes for each question, so forecasters could examine predicted results versus actual results to identify judgment errors and refine analytic techniques.

This particular study was valuable because it examined the cognitive process as a continuum rather than a solitary event, and explored a forecaster's potential to adapt and improve over time under multiple operant conditions. At the end of the experiment, researchers analyzed the data and found that forecasters who were held to some form of accountability consistently outperformed those who were not by a statistically significant margin. Individuals in the outcome condition outperformed process condition forecasters. Teaming amplified the accuracy of outcome forecasters, such that it was "about twice as effective at improving accuracy as process accountability, when compared to a no

accountability baseline."[151] Rather than regressing, the effect actually increased over time, and the accuracy of outcome accountable teams gradually improved throughout the duration of the study because forecasters learned and operationalized new strategies to leverage their teammates as a means of improving individual performance.

C. FORECASTER TRAINING

1. Less Wrong 101

There is a wealth of research dedicated to judgment and decision-making that collectively points to the frequency with which humans reach avoidably erroneous conclusions because of flawed probabilistic reasoning, cognitive bias, and irrational heuristics.[152] Given the criticality of accurate estimations to every facet of life, it is not surprising that the literature exploring strategies to improve the quality of human judgments via training related interventions is likewise abundant.[153] However, according to Welton Chang et al. in "Developing Expert Political Judgment: The Impact of Training and Practice on Judgmental Accuracy," their review of the literature associated with such strategies encountered several of the same shortcomings found in the accountability studies referenced previously.[154] The prevailing JDM training studies focused on isolated, solitary events, and typically failed to examine a forecaster's capacity to benefit from things like training, repetition, and feedback over a continuum. They likewise failed to assess whether

---

[151] Welton et al., 618.

[152] Maya Bar-Hillel, "The Base-Rate Fallacy in Probability Judgments," *Acta Psychologica* 44, no. 3 (May 1, 1980): 211–33, https://doi.org/10.1016/0001-6918(80)90046-3; Sarah Lichtenstein et al., "Judged Frequency of Lethal Events," *Journal of Experimental Psychology: Human Learning and Memory* 4 (November 1, 1978): 551–78, https://doi.org/10.1037/0278-7393.4.6.551; Paul Slovic and Baruch Fischhoff, "On the Psychology of Experimental Surprises," *Journal of Experimental Psychology: Human Perception and Performance* 3, no. 4 (1977): 544–51, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.365.7695&rep=rep1&type=pdf.

[153] Thomas R. Stewart, "Improving Reliability of Judgmental Forecasts," in *Principles of Forecasting*, ed. J. Scott Armstrong, vol. 30 (Boston: Springer U.S., 2001), 81–106, https://doi.org/10.1007/978-0-306-47630-3_5; Efraim Fischbein and Avikam Gazit, "Does the Teaching of Probability Improve Probabilistic Intuitions?: An Exploratory Research Study," *Educational Studies in Mathematics* 15, no. 1 (February 1984): 1–24, https://doi.org/10.1007/BF00380436; Baruch Fischhoff and Maya Bar-Hillel, "Focusing Techniques: A Shortcut to Improving Probability Judgments?," *Organizational Behavior and Human Performance* 34, no. 2 (October 1984): 175–94, https://doi.org/10.1016/0030-5073(84)90002-3.

[154] Welton Chang et al., "Developing Expert Political Judgment: The Impact of Training and Practice on Judgmental Accuracy in Geopolitical Forecasting Tournaments," *Judgment and Decision Making* 11, no. 5 (September 2016): 510, http://journal.sjdm.org/16/16511/jdm16511.pdf.

any improvements in subject performance were sustainable, or simply short-lived performance bumps subject to inevitable regression. Study subjects were also noted to lack any meaningful motivation to improve their performance, and the intervention(s) used by researchers in any given study tended to be tailored to mitigate a solitary heuristic defect rather than exploring the possibility to attack the problem holistically. GJP was an opportunity to test the hypothesis that improvements to judgmental accuracy resulting from training, feedback, and practice could be not only sustainable but also generalizable across a variety of judgmental challenges.[155]

GJP researchers publicly recruited a "highly diverse cross-section of the population" to serve as volunteer forecasters in the four year IARPA tournament, and randomly assigned them to training or control conditions.[156] At the beginning of each year, members of the training condition participated in a simple online instruction module designed to last no more than one hour. While the content of the training evolved somewhat over the course of the study, it remained a collection of best practice guidelines and strategies designed to help forecasters increase the accuracy of their predictions by avoiding common pitfalls, learning from errors, and maximizing the resources at their disposal.[157] Modules on probabilistic reasoning were accompanied by scenarios to reinforce key points. All of the information was supported with illustrative graphics, and each module concluded with knowledge checks or a brief quiz to allow students to assess their understanding.[158] Through a series of short lessons, forecasters learned that they could improve the quality of their thinking about uncertain outcomes by abandoning quick, intuitive judgments and instead adopting a more deliberate cognitive style informed by an understanding of probability, as well as the need to consider disconfirming possibilities and constantly update beliefs based on new information.

---

[155] Chang et al., 510.

[156] Chang et al., 510, 512.

[157] Philip Tetlock, "Edge Master Class 2015: A Short Course in Superforecasting, Class II," Edge, August 24, 2015, 4–5, https://www.edge.org/conversation/philip_tetlock-edge-master-class-2015-a-short-course-in-superforecasting-class-ii.

[158] Barbara Mellers et al., "Psychological Strategies for Winning a Geopolitical Forecasting Tournament," *Psychological Science* 25, no. 5 (May 2014): 2, https://doi.org/10.1177/0956797614524255.

2. Belief Updating

> Cable from Churchill to Lord Keynes: Am coming around to your point of view. Keynes reply: Sorry to hear it. Have started to change my mind.[159]
>
> —Author Noel Busch

The belief updating concept featured heavily throughout all forecaster training and originated from a theory developed by Thomas Bayes, a Presbyterian minister and statistician from London, whose grand contribution to probabilistic reasoning was released following his death in 1761.[160] Bayes' theorem incorporates multiple foundational concepts to improve probabilistic reasoning when the available data is scarce or inconclusive. The core concept involves starting with an initial estimate based upon an outside perspective of the problem space and gradually working inward, making incremental estimate updates at each interval based upon the value of the prior estimate, in light of the new evidence.[161] As an example, imagine a person challenged with guessing whether an obviously introverted student observed walking across the lawn of a major university is a member of the business school, or a math Ph.D. candidate.[162] An intuitive forecaster might reflexively reason that mathematics students are far more likely to be introverts when compared to business students and confidently assign a high probability that the student belongs to the math program. By contrast, a Bayesian trained forecaster starts by considering the outside perspective of the problem first, often in the form of historic frequency or total population. How many business school students are there in comparison to math Ph.D. candidates? Business school students are far more prevalent, so a reasonable estimate might be ten to one, thus there is a 90% likelihood that the observed student is enrolled in the business program. What portion of each population is introverted?

---

[159] Noel Busch, "Close-Up: Lord Keynes," *Life*, September 17, 1945, 122, https://books.google.com/books?id=t0kEAAAAMBAJ&q=%22a+cable%22#v=snippet&q=%22a%20cable%22&f=false.

[160] "Who Was Thomas Bayes?," Duke Today, accessed August 13, 2018, https://today.duke.edu/2012/11/bayes.

[161] "Bayes' Rule," University of British Columbia, accessed August 13, 2018, https://www.cs.ubc.ca/~murphyk/Bayes/bayesrule.html.

[162] "A Visual Guide to Bayesian Thinking," YouTube video, 11:24, posted by Julia Galef, July 16, 2015, https://www.youtube.com/watch?v=BrK7X_XlGB8.

One might assign a 75% likelihood for math and 15% for business, so based solely on that estimate, it is five times more likely that the correct answer is math. However, in considering the anterior estimate (1:10) in light of the posterior (75:15), the final ratio is 1:2, making it twice as likely that the shy student belongs to the business school, and that the intuitively attractive answer is ultimately incorrect.

A Bayesian approach to cognitive reasoning can provide utility across a broad range of complex problem spaces because it encourages forecasters to consider base rates, discard preconceptions, break seemingly intractable problems into smaller constituent parts, and think probabilistically. Nuclear physicist and Nobel laureate Enrico Fermi was famously fond of challenging students to use similar principles to estimate answers to superficially unanswerable questions using scarce supporting data.[163] One frequently cited example of a puzzle employed by Fermi is how many piano tuners are there in Chicago?[164] In lieu of a *go with your gut* answer, a GJP trained forecaster might begin by guessing the total population of Chicago, perhaps 2.5 million. In general, how many people in the United States actually have a piano? A blind guess might be one in one hundred, but if the number of institutions (universities, schools, music venues, etc.) owning pianos were also factored in, then maybe the ratio becomes two in one hundred. That would lead the forecaster to guess that there are about 50,000 pianos in Chicago. Some pianos are undoubtedly tuned more frequently than others, but on average, a neophyte might imagine that once per year is a reasonable frequency. Most people work 40 hours per week and spend two weeks per year on vacation, so an average piano tuner probably works 2,000 hours per year. Factoring in travel time to various job sites, and other administrative tasks, might decrease that number by 20%, bringing the total to 1,600 hours. How long does it take a piano tuner to tune a piano? If the forecaster guessed two hours, then a total of 100,000 tuning hours per year would be required to tune the 50,000 pianos estimated to be in Chicago. Dividing the total tuning hours required per year, by the number of hours a single tuner can devote to

[163] Andrea Peter-Koop, "Fermi Problems in Primary Mathematics Classrooms," *Australian Primary Mathematics Classroom Institute of Education Sciences* 10, no. 1 (2005): 1–5, https://files.eric.ed.gov/fulltext/EJ793997.pdf; Seth Shostak, "What Scientific Term or Concept Ought to Be More Widely Known? Fermi Problems," Edge, accessed August 15, 2018, https://www.edge.org/response-detail/27055.

[164] Tetlock and Gardner, *Superforecasting*, 110–14.

the task, produces a rough estimate that there are 63 piano tuners in Chicago. While it is unlikely that 63 is ultimately the correct response to the question, it is probably more accurate than a purely intuitive guess, and illustrates one effective strategy taught by GJP researchers to improve the accuracy of estimates associated with ambiguous or complex problems about which little is known. That style of outside-in thinking can provide forecasters with a course initial estimate, which can be further refined with increasing granularity as new evidence becomes available.

3. Sandbox Socialization: Learning How to Play Well with Others

In order to extract the maximum benefit from teams within the tournament, forecasters were explicitly educated about the advantages and disadvantages of working in groups. One of the dangers of sliding into groupthink is the tendency for teams to seek consensus, or subconsciously pilot themselves towards the least objectionable solution, rather than the most accurate result, so forecasters were advised to "never stop doubting" and to be "cooperative but not deferential" towards teammates.[165] Agreement within a group does not mean a particular judgment is correct, and disagreement is not an indication that something is wrong.[166] Indeed, disagreement can be beneficial as a means of expanding understanding and testing the validity of dissonant perspectives to refine judgment. Disagreement can alternatively produce acrimony and dysfunction if, for example, factions form behind clashing personalities, generating excessive heat and insufficient light.[167] Forecasters were trained to recognize indications of groupthink as well as specific interventions to guard against its insurgence within the team. GJP researchers taught forecasters how to "disagree without being disagreeable" and to actively use *constructive confrontation*, a term coined by former Intel CEO Andrew Grove, as a means of challenging assertions by disassembling arguments and asking precise questions to elicit

[165] Tetlock and Gardner, 199.

[166] Tetlock and Gardner, 199.

[167] Friedman, *Thank You for Being Late*, loc. 186.

detailed responses.[168] Forecasters also learned how to use teammates to assist with post-mortem analysis throughout the tournament season.[169] A post-mortem analysis takes place after a prediction has been submitted and the Brier score has been calculated as a means of evaluating the quality of the judgment that produced the prediction, in order to identify errors or opportunities for improvement.[170] Post-mortem analysis can illuminate "whether forecasters coalesced around a bad anchor, framed the problem poorly, overlooked an important insight, or failed to engage (or even muzzled) team members with dissenting views. Likewise, they can highlight the process steps that led to good forecasts and thereby provide…best practices for improving predictions."[171] Post-mortem analysis is valuable because, when considered in a vacuum, a good Brier score by itself does not necessarily indicate sound reasoning. For example, a parent who bet their child's entire college fund on a single game of roulette and won on the spin demonstrated good luck, not good judgment. Likewise, a poor Brier score might not necessarily indicate flawed thinking. A forecaster's prediction regarding anticipated growth within a specific regional economic sector might have produced an excellent Brier score had an unforeseeable *Black Swan* natural disaster not intervened.[172]

Red teaming and pre-mortem analysis both serve a similar function within teams but take place before a final prediction is submitted. Red teaming invites several forecasters within a group to form a devil's advocate brigade. Red Teams then attack a teammate's belief or line of reasoning in order to probe for weakness or identify flaws, thereby creating opportunities for improvement. In medicine, a pre-mortem analysis challenges a physician

---

[168] James Aisner, "Andy Grove: A Biographer's Tale," Harvard Business School, Working Knowledge, November 9, 2006, https://hbswk.hbs.edu/item/andy-grove-a-biographers-tale; Monica Worline and Dennis Matthies, *Stanford Learning Lab Learning Careers Project: A Self-Coaching Focus* (Stanford, CA: Stanford Center for Innovative Learning, n.d.), 3–5, http://scil.stanford.edu/research/learningcareers/documents/selfcoach1.pdf; Tetlock and Gardner, *Superforecasting*, 199.

[169] Chang et al., "Developing Expert Political Judgment," 514.

[170] Paul Schoemaker and Philip Tetlock, "Superforecasting: How to Upgrade Your Company's Judgment," *Harvard Business Review* 94 (May 2016): 8, http://mena-speakers.com/wp-content/uploads/2016/12/2016-hbr-final-final-version.pdf.

[171] Schoemaker and Tetlock, 13.

[172] Nassim Nicholas Taleb, *The Black Swan: The Impact of the Highly Improbable*, 2nd ed., Random House Trade pkb ed. (New York: Random House Trade Paperbacks, 2010), sec. Prologue.

to assume that a given diagnosis was wrong and resulted in the death of the patient. The objective is to gaze down at the imagined corpse on the slab and work backwards from the time of death to identify plausible scenarios for the patient's demise. GJP forecasters were shown how to use that technique within their own groups. Assume a prediction about a future event is wrong and the opposite comes to pass, then imagine the likely reasons that might explain the unanticipated outcome, as a means of evaluating the soundness of the original judgment. It is worth noting here that while the intractable singlemindedness observed in hedgehogs makes them comparatively poor forecasters, they are exceptionally adept at telling persuasive stories and asking good questions, which makes them valuable teammates for foxes wishing to test a hypothesis. This feature speaks further to the virtue of cognitive diversity within groups.

4. Hic Svnt Dracones

The root causes of many judgmental errors can be traced back to failures to understand and account for the role of probability in complex problem spaces, in concert with pervasive cognitive biases and flawed mental shortcuts used to span the gaps between known and unknown.[173] Humans have an inherent affinity for patterns coupled with an innate capacity to rapidly manufacture reason in response to uncertainty or ambiguity.[174] These evolutionary traits are both a blessing and a curse. A hunter's ability to instantly identify, interpret, and respond to nearly imperceptible environmental signals is vital. Rustling leaves may portend predator or prey, and the quality of the outcome is dictated by the speed with which the hunter can recall a similar experience to execute a solution. Conversely, forecasters who attempt to guess likely future outcomes of complex, multivariate events using reflexive, intuitive judgments without pausing to consider historic frequency or fundamental statistical principles are less likely to produce accurate predictions.

---

[173] Kahneman and Tversky, "On the Reality of Cognitive Illusions," 582–83, 589.

[174] Michael Shermer, "Patternicity: Finding Meaningful Patterns in Meaningless Noise," *Scientific American* 299, no. 48 (2008), https://www.scientificamerican.com/article/patternicity-finding-meaningful-patterns/; Jennifer A. Whitson and Adam D. Galinsky, "Lacking Control Increases Illusory Pattern Perception," *Science* 322, no. 5898 (October 3, 2008): 115–17, https://doi.org/10.1126/science.1159845.

The problem is well illustrated by Kahneman's presentation of U.S. cancer rates in *Thinking, Fast and Slow*.[175] Imagine a group of people asked to provide plausible explanations for the fact that among the 3,141 counties in the United States, those with the highest rate of kidney cancer are in small, rural, Southern communities. Depending upon a respondent's particular worldview, accusatory fingers might be pointed at a broad range of culprits. People in small rural communities are poor and poor people make bad health choices. They smoke and drink alcohol more than their more sophisticated urban neighbors. They cannot afford adequate healthcare. They tend to be superstitious and distrust the science of modern medicine. The U.S. healthcare system does not provide the same level of screening and care to poor people. Rural residents eat fast food in lieu of organic diets. Governments dump toxic industrial waste in rural communities. Rural communities offer fewer employment opportunities, so residents are forced to take hazardous jobs that expose them to harsh, potentially carcinogenic, environments. Poor people are genetically predisposed to X, Y, Z. It is related to the opioid crisis, because opioid abuse is more pronounced in small, rural communities?

The true culprit is far more pedestrian. Observed kidney cancer rates are higher in rural counties because they have small populations. Small numbers equal extreme results. The same phenomenon explains why the lowest rate of kidney cancer in the United States can also be observed in poor, rural counties. The failure to examine the problem from an outside statistical perspective, as one might consider sampling effects from representative marbles drawn from a jar, is only one common judgmental error represented by Kahneman's cancer question.[176] The failure to consider the historic frequency of kidney cancer throughout the entire population prior to responding is representative of *base-rate neglect*.[177] Seeking or assigning value to information based upon the extent to which it conforms to preconceived beliefs (i.e., poor people are superstitious) represents

[175] Kahneman, *Thinking, Fast and Slow*, loc. 109–10.

[176] Tversky and Kahneman, "Belief in the Law of Small Numbers."

[177] Ben Yagoda, "Your Lying Mind: The Cognitive Biases Tricking Your Brain," *The Atlantic*, September 2018, https://www.theatlantic.com/magazine/archive/2018/09/cognitive-bias/565775/.

confirmation bias.[178] Subconsciously estimating the frequency of an event or the likelihood of culpability based upon the ease with which a representative example comes to mind (the opioid epidemic) is described as a *heuristic of availability*, which is often accompanied by the heuristic of *attribute substitution*, in which people attempt to answer a perceptually complex question by unwittingly substituting the response to a simpler question instead.[179] A default presumption that higher kidney cancer rates in rural counties is linked to human behavior rather than independent circumstances is representative of the *fundamental attribution error*.[180] According to Kahneman, these illusions of knowledge persist because "we pay more attention to the content of the messages than to information about their reliability, and as a result end up with a view of the world around us that is simpler and more coherent than the data justify. Jumping to conclusions is a safer sport in the world of our imagination that it is in reality."[181] In each of the GJP online training modules presented over the course of the study, researchers provided forecasters with simple statistical and probabilistic reasoning principles.[182] Forecasters also received information related to specific cognitive biases commonly found to impair judgmental accuracy, as well as recommended strategies to mitigate their impact.[183]

5. Observed Training Outcomes

In contrast to notable research studies dedicated to measuring the impact of various JDM training interventions published prior to the IARPA tournament, the GJP training study considered interactive effects beyond training, such as practice, feedback, and

---

[178] Yeosun Yoon, Gülen Sarial-Abi, and Zeynep Gürhan-Canli, "Effect of Regulatory Focus on Selective Information Processing," *Journal of Consumer Research* 39, no. 1 (June 1, 2012): 93–110, https://doi.org/10.1086/661935.

[179] Daniel Kahneman, "Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice," in *Les Prix Nobel: The Nobel Prizes 2002*, ed. Tore Frangsmyr (Stockholm: Nobel Foundation, 2003), 465–74, https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2002/kahnemann-lecture.pdf.

[180] Philip E. Tetlock, "Accountability: A Social Check on the Fundamental Attribution Error," *Social Psychology Quarterly* 48, no. 3 (1985): 227–36, https://doi.org/10.2307/3033683.

[181] Kahneman, *Thinking, Fast and Slow*, loc. 113.

[182] Chang et al., "Developing Expert Political Judgment," 513–14.

[183] Mellers et al., "Improving the Accuracy of Geopolitical Risk Assessments," 7.

experience, to measure their comprehensive influence over time. Before the beginning of the first year of the IARPA tournament (September 2011 through April 2012), forecasters randomly assigned to the training condition received a brief (~ one hour) block of on-line instruction designed to improve probabilistic estimates.[184] Over the course of the following nine months, tournament officials tasked all of the GJP study participants (trained condition and non-trained control) with answering approximately 150 questions related to a topically diverse array of complex future geopolitical events. This provided members of the trained condition with an opportunity to immediately operationalize their training. Participants received constant feedback throughout the year in the form of Brier scores, which allowed them to assess and refine their analytic technique. The same process was repeated for each of the three remaining tournament years, which provided forecasters with significant experience and opportunities for learning.[185] At the end of the fourth year, the IARPA tournament was closed. Welton Chang et al. analyzed the data and published the results.

At the end of the first tournament year, forecasters randomly assigned to the trained condition outperformed untrained control members by 10%.[186] At the end of the second year, the variation between the two groups rose to 12%.[187] The effect was statistically significant throughout all four years of the tournament without regression. Overall, training accounted for ~ 10% of the accuracy variance observed between condition and control forecasters.[188] To understand the deep significance of that superficially modest result, one might conjure the image of a university wherein every student's GPA increased by one letter grade because he or she watched a 60-minute webinar during orientation. Now imagine that same result, occurring to every student, every year, for the following three years.

---

[184] Mellers et al., "Psychological Strategies for Winning a Geopolitical Forecasting Tournament."

[185] Schoemaker and Tetlock, "Superforecasting," 6.

[186] Chang et al., "Developing Expert Political Judgment," 514.

[187] Chang et al., 514.

[188] Barbara Mellers et al., "The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics," *Journal of Experimental Psychology: Applied* 21, no. 1 (2015): 97, https://doi.org/10.1037/xap0000040.

D. FORECASTER TRACKING: KEEPING SCORE

> Despite the money it costs, despite its importance in government decision-making, the quality of so much forecasting remains untested, and therefore unknown. Forecasters never learn what they're getting right and wrong. They never adjust what they're doing to make their forecasts better.[189]
>
> —Authors Philip Tetlock and Dan Gardner

1. The Failure to Measure

In EPJ, Tetlock surmised that one of the reasons experts, analysts, and pundits were able to make repeated, unambiguously inaccurate predictions about consequential future geopolitical events with impunity was because they were never called to task. No one was systematically tracking and reporting whether predicted events manifested into actual events. Tetlock observed, "Although there is nothing odd about experts playing prominent roles in debates, it is odd to keep score, to track expert performance against explicit benchmarks of accuracy and rigor."[190] This vacuum leaves the consumers of expert predictions free to assess the quality of an expert based upon purely subjective criteria, such as their ability to tell convincing stories, prestigious titles, academic clout, popularity, or the degree to which a narrative compliments or conforms with a consumer's pre-conceived worldview.[191] This likewise frees experts from the otherwise exhaustively rigorous task of producing objectively accurate predictions of uncertain future events, in exchange for the comparatively simpler task of mollifying a constituent base of forecast consumers.[192]

---

[189] Philip Tetlock and Dan Gardner, "We Can Learn to Predict Future Events," *The Telegraph*, October 30, 2015, 1, http://www.telegraph.co.uk/news/uknews/defence/11965831/We-can-learn-to-predict-future-events.html.

[190] Tetlock, *Expert Political Judgment*, 1.

[191] Tetlock and Gardner, *Superforecasting*, 5.

[192] "Edge Master Class 2015—Philip Tetlock: A Short Course in Superforecasting," Edge, accessed October 11, 2017, https://www.edge.org/event/edge-master-class-2015-philip-tetlock-a-short-course-in-superforecasting.

Ted Knight: What did you shoot?
Chevy Chase: Oh, I don't keep score.
Ted Knight: Then how do you compare yourself to other golfers?
Chevy Chase: By height.[193]

Dr. John McCreary, a prominent retired senior analyst at the Defense Intelligence Agency (DIA), recipient of the Presidential Rank Award, and editor of *NightWatch*, finds the same scoring void persists within the U.S. intelligence community (IC).[194] According to McCreary, the failure to track and record the quality of analytic reports has resulted in accuracy rates for critical U.S. intelligence estimates that are "borderline criminal."[195] In a 2017 e-mail to NPS professor and former Chief of Naval Intelligence Robert Simeral, McCreary reported that by his own measurements over a period beginning in 2006, the judgmental accuracy of forecasts prepared for the Joint Chiefs of Staff was 48%. Within that context, replacing the DIA with a dull quarter would be a defensible strategy.

2. Delusions of Measurement

> The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge.[196]
>
> —Author K. S. Somesh

The failure to quantitatively correlate anticipated outcomes with actual outcomes and adjust beliefs accordingly is certainly not limited to domains of intelligence, policy, or

---

[193] *Caddyshack*, directed by Harold Ramis (1980; Los Angeles, CA: Orion Pictures, 1980), https://www.imdb.com/title/tt0080487/; Charles J. Wheelan, *Naked Statistics: Stripping the Dread from the Data*, 1. publ. as a Norton pbk (New York: Norton, 2014), 44.

[194] "John F. McCreary: Executive Profile & Biography," Bloomberg, accessed July 13, 2018, https://www.bloomberg.com/research/stocks/private/person.asp?personId=267776426&privcapId=733543&previousCapId=733543&previousTitle=Kforce%20Government%20Solutions,%20Inc.

[195] "FAO Asia In-Residence Course 21 June–2 July 2010—Intelligence Analysis and Professionalism Mr. John McCreary," Naval Postgraduate School, video, July 27, 2010, http://web.nps.edu/Video/Portal/Video.aspx?enc=MPCfFMbHCWexu6JsvpD%2FxQDYA8NNCnml; John McCreary, "Professionalism in Analysis" (lecture, Naval Postgraduate School, CA, December 23, 2009); Robert Simeral, email message to author, March 15, 2017.

[196] K. S. Somesh, "The Greatest Enemy of Knowledge Is Not Ignorance, It Is the Illusion of Knowledge," *Medium* (blog), March 8, 2018, https://medium.com/@ks.somesh2016/the-greatest-enemy-of-knowledge-is-not-ignorance-it-is-the-illusion-of-knowledge-5c0dd1dcca7e quoting Neil Boorstin.

public discourse. In *A Failure of Risk Management*, Douglas Hubbard, author and inventor of applied information economics, identified the catastrophic results of a broad range of high consequence events (natural disasters, aviation incidents, critical infrastructure failures, industrial accidents, economic calamities, etc.), which cluster around a single fault point; an organizational preference for perception over precision.[197] In the preface of the book, Hubbard succinctly summarized the problem by saying "risk management based on actual measurements of risks is not the predominant approach of most industries."[198] *Risk*, according to Hubbard, is simply the likelihood that something bad could happen in the future, and *management* is the use of available resources to achieve a desired result.[199] Risk management therefore, as described by Hubbard, is "the identification, assessment, and prioritization of risks followed by coordinated and economical application of resources to minimize, monitor, and control the probability and/or impact of unfortunate events," or alternatively, "being smart about taking chances."[200]

When Hubbard wrote about the *failure* of risk management, he was not only referring to the absence of any scientifically valid process for identifying and mitigating future threats. A portion of the book is dedicated to a more nuanced problem in which organizations "believe they have adopted an effective risk management method and are unaware that they haven't improved their situation one iota."[201] Hubbard is referring to instances in which otherwise sophisticated groups overconfidently rely upon pseudoscientific methods of risk estimation for which there is "almost no experimentally verifiable evidence" that they actually work.[202] Hubbard anecdotally recounted a national pharmaceutical conference he attended in 2007, in the wake of multiple high profile public safety incidents resulting from chemical production outsourcing to China.[203] The attendees

---

[197] Hubbard, *The Failure of Risk Management*, 1–8.

[198] Hubbard, XI.

[199] Hubbard, 8–9.

[200] Hubbard, 10.

[201] Hubbard, *The Failure of Risk Management*, 16.

[202] Hubbard, 17.

[203] Hubbard, 11–12.

were all experienced and well-credentialed scientists, chemists, and engineers. The agenda was comprised of technical presentations detailing specific chemical design, production, and packaging methodologies, all of which were accompanied by complex mathematical models. Skeptic attendees actively questioned or challenged every presenter's assertion in a manner befitting professional scientists. The character of the convention hall changed dramatically when the topic transitioned to a new methodology for evaluating the risks associated with outsourcing drug manufacturing to China. The model was entirely based on a subjectively weighted scoring method utilizing a 1 to 5 scale across multiple categories. The weights, as well as the scores contained therein, were the sole product of opinions originating from a small, designated group within the company. Drugs that did not exceed a pre-specified threshold were deemed safe for outsourcing.

At the end of a one-sided presentation, the sole question from the floor was offered by Hubbard, who laconically asked, "How do you know it works?"[204] The precipitated silence was deafening, and did not abate. In contrasting the fervent academic pitch that permeated the hall prior to the outsourcing presentation with the viscous snake oil salesmanship thereafter, Hubbard sardonically remarked that the assembled mass of analytic aces had tacitly endorsed "an approach with no more scientific rigor behind it than an ancient shaman reading goat entrails" and further observed, "While the lack of such rigor would be considered negligent in most of their work, it was acceptable to use a risk assessment method with no scientific backing at all."[205] In the realm of forecasting future outcomes, overt negligence might take a back seat to comforting delusions.

[204] Hubbard, 13.

[205] Hubbard, 14.

3. Tracing the GJP Data Loop

Data! data! data! he cried impatiently. I can't make bricks without clay.[206]

—British writer Arthur Doyle

GJP's goal in the IARPA tournament was to produce the most accurate predictions of complex future geopolitical events as was humanly possible. Over the course of four years, Tetlock's crowdsourced team provided more than one million time series forecasts in response to approximately 500 questions posed by IARPA officials.[207] The first question posed in the first year of the tournament was "Will the Six-Party talks (among the United States, North Korea, South Korea, Russia, China, and Japan) formally resume in 2011?"[208] The question was published on September 1, 2011, at which time GJP forecasters could submit a prediction via an online website in the form of a binary yes/no response accompanied by a confidence interval (e.g., 63% Yes). As long as the question remained open, forecasters could update their predictions as often as they chose based on (for example) evolving beliefs or the acquisition of new information deemed pertinent from any unclassified source (teammate conversations, Google, etc.). With each prediction, forecasters were encouraged to use the website to capture personal notes in order to memorialize the reasoning or thought process that precipitated the estimate. This facilitated a more detailed post-mortem analysis and provided additional insight for researchers.

4. Aggregating and Averaging

In 1989, Dr. Robert Clemen published a comprehensive literature review of more than 200 academic papers dedicated to prediction, which found unanimous consensus amongst all of the contributing researchers that "combining multiple forecasts leads to

---

[206] Arthur Doyle, "The Adventure of the Copper Beeches," in *The Adventures of Sherlock Holmes*, 2017, loc. 3925 of 4318, Kindle, https://www.amazon.com/Adventures-Sherlock-Holmes-Arthur-Conan-ebook/dp/B06XPLKCSB/ref=tmm_kin_swatch_0?_encoding=UTF8&qid=&sr=.

[207] Good Judgment Project, "GJP Data," Harvard Dataverse, 2016, https://doi.org/10.7910/DVN/BPCDH5.

[208] Good Judgment Project, "GJP Data Year One," Harvard Dataverse, 2016, https://doi.org/10.7910/DVN/BPCDH5.

increased forecast accuracy… Furthermore, in many cases one can make dramatic performance improvements by simply averaging the forecasts."[209] Accordingly, beginning on the morning of September 2, 2011, and continuing every day that the question was open thereafter, Tetlock and his team aggregated and averaged the predictions submitted by GJP forecasters and distilled them into a single best guess (in virtually the same manner as Galton more than a century before).

5. Extremizing

This simple algorithmic manipulation was the first step in GJP's attempt to alchemically transmute lead into gold. The second step was a process dubbed *extremizing*. In a 2015 lecture to the *American Enterprise Institute*, Tetlock explained GJP's rationale for extremizing predictions by asking attendees to imagine a room full of U.S. intelligence experts convened by President Obama in 2011 to present their individual estimates of whether the person inside a walled compound in Abbottabad, Pakistan was Osama bin Laden. Suppose the reported confidence level from every person in the room was 65%. Based on that, would it be reasonable for the President to conclude that there was a 65% chance that the tall figure was bin Laden? According to Tetlock, it depends. If the analysts had reached their individual conclusions after considering the exact same intelligence, then 65% might represent the best possible guess. However, suppose that each contributor represented a different compartmentalized segment of the IC (human, signal, image, etc.) and had based their estimate exclusively upon the information available within their respective silo. Would 65% still represent the best possible guess? While it is impossible to know for certain, Tetlock suggested that the probability should be adjusted, or extremized, based upon the diversity of the inputs into the system. An analyst who reported a 65% likelihood based on the limited intelligence contained within his or her silo would almost certainly increase their level of confidence if the silos were destroyed and all of the information became available. Were that to happen, the President might reasonably conclude that the most accurate estimate is 70% or 75%, despite the fact that no individual advisor guessed above 65%. Tetlock surmised that the same phenomenon would apply to

[209] Clemen, "Combining Forecasts," 559.

GJP forecasters. In attempting to answer the Six-Party question, each forecaster could use any available unclassified source of information to produce a guess. As a result, it was impossible for one forecaster to know everything that all the other forecasters knew. If it were possible, then GJP's collective confidence would increase and guesses would be adjusted a little bit closer to 100% or 0%. Extremizing then, is a way to simulate how a group's estimate would react *if* all of the forecasters could pool and collectively understand all of the available information used by the entire GJP team. If the collective responses trend towards 100% or 0% when averaged, GJP simply nudged the final estimate a little more in proportion to the degree of the prevailing direction of the crowd (hence, forecasts that resulted in 70% yes when averaged are extremized more than forecast averages yielding 60% yes).

At midnight (EST) on December 31, 2011, the Six Party question was closed. The following morning, GJP aggregated and averaged all of the forecasts submitted by its members to produce a single prediction, and then extremized the result (representing GJP's official best guess), which was electronically transmitted to IARPA at 9:00 AM. On January 2, 2012, IARPA declared the correct answer to the question to be "no."

At this point, IARPA scored GJP's performance using a method developed in 1950 by statistician Glenn W. Brier.[210] Originally designed to allow weather forecasters to assess the accuracy of weather predictions, IARPA used Brier scoring because it accounts for a prediction's resolution (the event happened or did not happen) as well as calibration (the confidence interval assigned to whether a future event will happen or not). Simply put, a forecaster who predicts a 75 percent chance that a future event will take place will score much better than one who only placed the odds at 51%. A Brier score represents the degree of error in a forecast. To calculate a score, the forecasted probability of an event is divided by 100, so the result is between zero and one. The outcome of the event is then assigned either a 0 (did not happen) or a 1 (did happen). For each potential outcome, the differences

[210] Glenn W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review* 78, no. 1 (January 1, 1950): 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2; "Frequently Asked Questions (FAQ)," Good Judgment® Open, last updated June 16, 2017, https://www.gjopen.com/faq.

between the predicted outcome and the actual outcome are squared and then added together to obtain a score. For example, if a forecaster correctly assigns a 70% likelihood that an event will happen, the Brier score is .18, or the sum of $(1\text{-}0.7)^2 + (0\text{-}0.3)^2$. Brier scores can range between zero and two, with zero reflecting perfect clairvoyance.

Brier scores allowed researchers to begin meticulously tracking the performance of every GJP forecaster over time. This feedback facilitated post-mortem analysis and enabled learning. Researchers publically posted the scores in online leaderboards so forecasters could compare their performance against other GJP forecasters, and teams could benchmark themselves against other teams.[211] This facilitated recognition and competition, which is discussed in a subsequent section.

6. Weighting

Tracking and recording the performance of every GJP forecaster also facilitated the implementation of a third algorithmic manipulation, intended to work in concert with averaging and extremizing. As discussed in a preceding section, large groups of diverse forecasters are capable of producing accurate predictions because scraps of useful information are widely dispersed among group members, and inaccurate information is typically distributed randomly above and below the correct answer such that when averaged on a graph, good guesses gain resolution and bad guesses cancel one another out. While a group's prediction is frequently more accurate than the best guess of any individual member, such is not always the case. Certainly there are instances in which a single member beats the group; however, over time it is unlikely that the same individual will come out on top. In the long run, given the option between consistently betting on one person or one group, smart money wagered on the group will pay out with greater frequency than bets placed on one individual, thereby making groups the better gamble. However, GJP's terminal objective was not to produce better outcomes. Their goal was to produce the best possible outcome. Researchers surmised that while the same forecaster

[211] Barbara Mellers et al., "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions," *Perspectives on Psychological Science* 10, no. 3 (2015): 269, http://journals.sagepub.com/doi/abs/10.1177/1745691615577794.

was unlikely to consistently beat the group, performance tracking over time would reveal that the same sub-set of forecasters would perform at the top of the scale, representing a heretofore untapped resource ripe for exploitation to further improve GJP's terminal accuracy. Tetlock et al. reasoned that some forecasters are better than others, and over time, the cream should rise to the top. This notion was subsequently well supported by David Budescu and Eva Chen in a paper entitled "Identifying Expertise and Using It to Extract the Wisdom of the Crowds."[212] When the follow-on to the Six Party question was opened by IARPA, GJP proceeded in exactly the same fashion as it had before; however, this time the responses from the forecasters at the top of the leaderboard were artificially weighted based upon the quality of their prior performance. Those weights were incorporated into the aggregated average, which was then extremized to produce a final GJP prediction that was transmitted to IARPA once the question was closed.

7. Superteams

The process of weighting, aggregating, averaging, and extremizing predictions continued in response to every question posed by IARPA for the duration of the first tournament year. After the first season of the IARPA tournament was concluded, GJP examined the cumulative performance of all forecasters and observed that a core 2% minority of the total group, dubbed superforecasters, had indeed consistently outperformed the rest.[213] Researchers wanted to make further use of that feature as a means of extracting greater levels of accuracy from the system by re-grouping previously random teams into performance stratified teams. The prospect of doing so was not without risk.[214] On one hand, there was a significant body of literature centered around student performance which held that cohorts grouped by superior ability performed better than those grouped

[212] David V. Budescu and Eva Chen, "Identifying Expertise and Using It to Extract the Wisdom of the Crowds," *Management Science* 61, no. 2 (May 23, 2014): 37, http://pages.stern.nyu.edu/~eyoon/seminar/dbudescu/Paper.pdf.

[213] Tetlock and Gardner, *Superforecasting*, 201.

[214] Mellers et al., "Psychological Strategies for Winning a Geopolitical Forecasting Tournament," 2–3.

heterogeneously, largely owing to an accelerated learning effect.[215] Alternatively, making forecasters aware that they had been anointed with the "super" label could produce debilitating arrogance or overconfidence.[216] Researchers also looked for guidance beyond the established literature by informally asking recognized organizational experts for their opinions.[217] The responses were "flatly contradictory."[218] Ultimately, Tetlock decided to roll the dice and reorganized all of the GJP teams according to past performance. As a result, prior to the start of the second season of the IARPA tournament, five superteams of 12 forecasters each (representing the top 2% of performers from the previous year) were created. The resultant outcomes observed at the end of the second tournament year were wholly unanticipated.

If the performance levels produced by superforecasters at the end of the first year were the product of luck rather than skill, it would be reasonable to expect them to regress towards the mean in year two. They did not.[219] Tetlock would later liken the creation of superteams to a cognitive "steroid injection" for those forecasters who were already performing at the very top of the scale.[220] GJP found, "On average, when a forecaster did well enough in year 1 to become a Superforecaster, in year 2 that same person became 50% *more* accurate. An analysis following year three produced the same result."[221] Researchers naturally wanted to broaden their understanding of the superteam phenomenon and its

---

[215] Dennis Epple and Richard E. Romano, "Peer Effects in Education: A Survey of the Theory and Evidence," in *Handbook of Social Economics*, eds. Jess Benhabib, Alberto Bisin, and Matthew O. Jackson, vol. 1 (San Diego: North-Holland, 2011), 1053–1163, https://doi.org/10.1016/B978-0-444-53707-2.00003-7; Julian R. Betts and Jamie L. Shkolnik, "The Effects of Ability Grouping on Student Achievement and Resource Allocation in Secondary Schools," *Economics of Education Review* 19, no. 1 (February 2000): 1–15, https://doi.org/10.1016/S0272-7757(98)00044-2.

[216] Mellers et al., "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions," 269.

[217] Philip E. Tetlock et al., "Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate," *Current Directions in Psychological Science* 23, no. 4 (August 2014): 291, https://doi.org/10.1177/0963721414534257.

[218] Tetlock et al., 291.

[219] Barbara Mellers et al., "How Generalizable Is Good Judgement? A Multi-Task, Multi-Benchmark Study," *Judgement and Decision Making* 12, no. 4 (July 2017): 370, http://journal.sjdm.org/17/17408/jdm17408.pdf.

[220] Mellers et al., "Psychological Strategies for Winning a Geopolitical Forecasting Tournament," 8.

[221] Schoemaker and Tetlock, "Superforecasting," 205.

mediating influence upon superforecasters in the form of an enriched learning environment, and to that end they had a valuable tool at their disposal. GJP forecasters were loosely networked via a purpose-built online website that captured all of their interactions. Researchers were able to use that platform to construct queries measuring superforecaster behaviors prior to and after the creation of superteams, as well as queries comparing superteam behaviors with those observed in *top-teams* (one performance tier below superteams) and *all-other* teams. For example, at the end of the third year of the IARPA tournament, GJP found that:[222]

- First year superforecasters updated their beliefs twice as often as any other forecaster. The frequency doubled after they were pooled into superteams at the beginning of the second year, and the rate increased again in the third year.
- The average number of questions attempted by superforecasters in the first year increased by 30% following placement within a superteam.
- Between years two and three, superteams collected articles and performed Google keyword searches 400% more than other teams.
- Between years two and three, superteams posted approximately 500% more question-specific comments than all other teams.
- Comments posted by superteams between years two and three were 30% longer than those of all other teams.
- Between years two and three, superteams shared question-relevant information with teammates in the form of article links and file attachments 10 times more frequently than other teams.

[222] Mellers et al., "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions," 274, 277–78.

- Superteams asked their teammates twice as many questions as any other team, and were five times more likely to receive a response.

In summary, GJP found that pooling forecasters into performance stratified teams created enriched environments that drove performance gains. For newly minted superforecasters who were already performing at the top of the charts, this superteaming effect produced significantly increased demonstrations of task motivation and commitment. Superforecasters assigned to superteams demonstrated a greater motivation to cultivate new skills, as well as increased levels of internal engagement, and a greater willingness to share knowledge. Members of superteams were more likely to ask teammates for help, and more likely to receive it. Overall, they reflected a higher degree of perceived accountability to one another. Each of these features was ultimately determined to be predictors of increased accuracy based upon tournament Brier scores.[223]

Performance = Recognition = Engagement = Rigor = Learning = Performance

In addition to facilitating the weighting, extremizing, and teaming methodologies employed by GJP to win the IARPA tournament, performance tracking fueled an additional feature that contributed to their ultimate success, forecaster recognition. Within GJP, accuracy was the only true currency, making public leaderboards "radically meritocratic" sources of healthy competition, acclaim, and validation amongst forecaster peers.[224] The psychological impact of positive recognition was recently described by cognitive scientist turned professional poker player Annie Duke in her book, *Thinking in Bets*.[225] Duke, now a world champion, recalled herself as a young proselyte player having just joined an elite group of successful, highly skilled players whose purpose (much like a superteam) was to help one another develop improved judgmental accuracy within the poker realm. Recognition within the group, and approval from it, exerted a narcotic influence on the

[223] Mellers et al., 277.

[224] Philip Tetlock, "How to Win at Forecasting: A Conversation with Philip Tetlock," Edge, 9, accessed August 3, 2018, https://www.edge.org/conversation/philip_tetlock-how-to-win-at-forecasting; Mellers et al., "Improving the Accuracy of Geopolitical Risk Assessments," 5.

[225] Annie Duke, *Thinking in Bets: Making Smarter Decisions When You Don't Have All the Facts* (New York: Portfolio Penguin, 2018).

fledgling felt dweller. According to Duke, "I experienced firsthand the power of a group's approval to reshape individual thinking habits. I got my fix by *trying* to be the best credit-giver, the best mistake-admitter, and the best finder-of-mistakes-in-good-outcomes. The reward was their enthusiastic engagement and deep dives introducing me to the nuances of poker strategy. It was also rewarding to have these intelligent, successful players take my questions seriously and increasingly ask for *my* opinions."[226]

8. The Glory of GitHub

> There is something wonderfully human about the open-source community. At heart, it's driven by a deep human desire for collaboration and a deep human desire for recognition and affirmation of work well done—not financial reward. It is amazing how much value you can create with the words "Hey, what you added is really cool. Nice job. Way to go!" Millions of hours of free labor are being unlocked by tapping into people's innate desires to innovate, share, and be recognized for it.[227]
>
> —Writer Thomas Friedman

Recognition is credited with the meteoric rise of a massive online open source software development platform. Launched in 2008, *GitHub* was a digital space designed to facilitate crowdsourced collaboration on software design projects in which anyone could contribute to the process or benefit from the proceeds.[228] The passion for creative innovation and the tremendous effort exerted by programmers throughout the world to build products that no one could own was driven by the basic human need for positive peer recognition.[229]

---

[226] Duke, 133.

[227] Friedman, *Thank You for Being Late*, loc. 1127.

[228] "GitHub Features: The Right Tools for the Job," GitHub, accessed August 30, 2018, https://github.com/features.

[229] Laura Dabbish et al., "Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work—CSCW '12*, ACM 2012 conference (Seattle: WA: ACM Press, 2012), 1277, https://doi.org/10.1145/2145204.2145396.

New projects were born when someone uploaded a prototype design which any user could then experiment with, modify, expand, or improve. GitHub was completely transparent, such that every contribution was instantly attributable to its contributor, and all users could appraise the value of that specific contribution via a peer rating system. Good contributions were incorporated into the prototype, bad contributions were discarded, or alternatively portions of the prototype could be extracted and used to create wholly different projects. GitHub co-creator Chris Wanstrath described it as "a distributed version controlled system: anyone can contribute, and the community basically decides every day who has the best version… The best rises to the top by the social nature of collaboration—the same way books get rated by buyers on Amazon.com."[230] Peer reviews at every step generated praise, criticism, conversations, and debate, which produced "a virtuous cycle for the rapid learning and improvement of software programs that drives innovation faster and faster."[231] In speaking about the source of GitHub's success, Hewlett Packard President Meg Whitman simply said, "I am convinced that the world is driven by validation and that's what makes these communities so powerful. People are driven by their desire for others in the community to validate their work."[232]

9. Ripping Warez: Piracy in High C

In a relevant albeit unconventional vein, consider the feature that fanned the flames of the bonfire that was the illegal music trade that sparked to life in 1995. This same factor, ironically, would later reduce the towering pixelated pyre to a soggy pile of inert embers: recognition. As far back as the 1980s, *Warez* was a term ascribed to copyrighted digital media with copy protection barriers that had been cleverly circumvented to facilitate unlawful distribution, as well as the virtually connected network of people who produced and propagated the ill-got gains.[233] An offshoot of the Warez piracy enterprise dedicated

---

[230] Friedman, *Thank You for Being Late*, loc. 1066.

[231] Friedman, loc. 1034.

[232] Friedman, loc. 1134.

[233] Ard Huizing and Jan A. van der Wal, "Explaining the Rise and Fall of the Warez MP3 Scene: An Empirical Account from the Inside," *First Monday* 19, no. 10 (October 6, 2014): 1, http://journals.uic.edu/ojs/index.php/fm/article/view/5546.

exclusively to music eventually emerged following the rise of the internet and the creation of the ubiquitous MP3 file format. Colloquially and collectively referred to as the *MP3 scene*, this subset quickly self-organized into distinct, socially interconnected release groups, with members located throughout the world.

The end-to-end process of illicitly distributing contraband concertos was both labor and time intensive, but more critically, demanded tremendous technical skill and creativity. A notable characteristic of the MP3 scene was that, like GitHub, it was a non-monetary economy.[234] The only currency of commerce was positive social performance feedback.[235] Individual and team performance metrics were measured and recorded based upon "the ability to continuously release new files before anyone else, preferably before the official release." These accomplishments brought about positive social recognition and enhanced reputation for the individuals and teams responsible.[236] Task specific roles within teams required specialization, such that they were compartmentalized and interdependent, making individual performance and group performance inextricably linked. This generated a "self-motivating virtuous cycle of intrinsic flow experiences, extrinsic rewards, and a 'copyfight' culture stimulating feelings of spontaneous pleasure and a recurrent desire for more that powered participants' passion for music and knowledge, and their emotional attachment to and identification with the scene, which, on their turn, reinforced members' intrinsic and extrinsic motives for participation."[237] The recording industry responded to the revenue loss by initiating legal countermeasures and incorporating improved digital protections for their software. These required greater creativity, skill, and effort to overcome, which resulted in increased opportunities for acclaim, which improved pirate performance. In the face of steadily increasing (and well-funded) industry resistance, the recognition induced *flow state* initially described by Mihaly Csikszentmihalyi in his book, *Flow: The Psychology of Optimal Experience*, crescendoed the MP3 scene to a point in

---

[234] David McCandless, "Warez Wars," Wired, April 1, 1997, https://www.wired.com/1997/04/ff-warez/.

[235] Alf Rehn, "The Politics of Contraband," *The Journal of Socio-Economics* 33, no. 3 (July 2004): 359–74, https://doi.org/10.1016/j.socec.2003.12.027.

[236] Huizing and Wal, "Explaining the Rise and Fall of the Warez MP3 Scene," 5.

[237] Huizing and Wal, 6.

2005 in which nearly 25,000 unique copyrighted songs were being released to the underground for free *per month*.[238] Soon thereafter, the entire MP3 scene experienced a precipitous decline in productivity from that high-water mark which was brought about by advances in hardware and software.[239] The process of producing counterfeit music became faster, more efficient, and easier. This de-skilling effect transformed praiseworthy into prosaic, and the engine that drove the industry sputtered, coughed, and died.

10. Vivisecting Visionaries: The Cognitive Anatomy of a Superforecaster

One final feature of performance tracking that merits consideration is that it provided researchers with the means to ask and answer a nagging question, are Superforecasters born or made? GJP forecasters completed a battery of standardized tests designed to measure various facets of cognitive ability, motivation, and style.[240] Researchers used univariate regression analysis of the independent variables produced by testing results, correlated with a dependent variable represented by each subject's mean Brier score, to determine the extent to which each of the various quantifiable differences between individual forecasters contributed to the variance observed in predictive accuracy.[241] All told, the answer to the nature/nurture question was a resounding: yes.

The following variables were found to be consistent correlates for judgmental accuracy:[242]

- Superforecasters are intelligent. Cognitive ability moderated accuracy and was a consistently significant performance predictor throughout the

[238] Huizing and Wal, 3.

[239] Huizing and Wal, 8.

[240] Tetlock and Gardner, *Superforecasting*, 107–10.

[241] Chang et al., "Developing Expert Political Judgment," 518–19.

[242] Chang et al., 511–22; Mellers et al., "Psychological Strategies for Winning a Geopolitical Forecasting Tournament," 8–9; Mellers et al., "The Psychology of Intelligence Analysis," 91, 96–101; Mellers et al., "How Generalizable Is Good Judgement? A Multi-Task, Multi-Benchmark Study," 370–370; Mellers et al., "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions," 272–80; Mellers et al., "Improving the Accuracy of Geopolitical Risk Assessments," 5–7.

tournament. These results are consistent with the collective output from decades of research.[243]

- Superforecasters are numerate. This is not to say that all superforecasters are skilled mathematicians, only that they are generally more comfortable with thinking in numbers than their less accurate peers coupled with an affinity for probability.

- Superforecasters are actively open-minded. They view beliefs as "hypothesis to be tested, not treasures to be guarded."[244] They search for contradictory evidence and are tolerant of ambiguity. They do not declare ideological allegiance to any tenet or ethos. Accuracy alone is sacred. All other beliefs are subject to revision based upon available evidence. The frequency with which forecasts were updated was a consistent predictor for accuracy throughout the tournament.

- Superforecasters have a greater need for intellectual challenges. They are drawn to mental puzzles and problem solving.

- Superforecasters are competitive and driven by competition for status and recognition, yet actively seek opportunities to share and collaborate, because it increases their odds of winning.

- Superforecasters have a secular worldview and do not believe in pre-deterministic perspectives that seek to attribute outcomes to fate or the supernatural. They understand and accept the role of randomness in the world.

---

[243] Frank L. Schmidt and John E. Hunter, "The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings," *Psychological Bulletin* 124, no. 2 (1998): 265, http://psycnet.apa.org/psycinfo/1998-10661-006; Frank L. Schmidt and John Hunter, "General Mental Ability in the World of Work: Occupational Attainment and Job Performance," *Journal of Personality and Social Psychology* 86, no. 1 (2004): 162–73, https://doi.org/10.1037/0022-3514.86.1.162.

[244] Tetlock and Gardner, *Superforecasting*, 191.

One consistently observed trait in every superforecaster was the tenacious pursuit of self-improvement. None perceived judgment to be an innate, immutable commodity, but rather a skill that could be improved upon over time with sufficient diligence and practice; a state of mind referred to by Tetlock as "perpetual beta."[245] Are superforecasters born or made? The data supports both conclusions. Some forecasters are unquestionably much better than others. Innate dispositional characteristics like intelligence are vital and absolutely matter. That said, environmental factors, such as group diversity, performance stratified teams, training, practice, accountability, and recognition are transformative. Ignoring one aspect in deference to the other is perilous. Anyone who believes otherwise has exercised "demonstrably poor judgment."[246]

E. RECONSTRUCTING TETLOCK: IARPA RESULTS

Once combined, the cumulative effects of teaming, training, and tracking produced remarkable outcomes.[247] At the end of the first tournament year, GJP beat the IARPA control group by 60%. After year two, that number rose to 78%. At the end of the second year, GJP had also bested the other four university teams by 30–70%, which prompted IARPA to drop them from the tournament.[248] GJP's winning margins increased for each of the two remaining years of the tournament and was ultimately declared the victor. After more than one million predictions, GJP was found to be on the right side of maybe more than 86% of the time.[249] Most astoundingly, a journalist later leaked a secret report that revealed that GJP "nobodies" armed only with Google beat an undisclosed shadow team of DNI professionals armed with classified intelligence by 30%.[250]

---

[245] Tetlock and Gardner, 190, 192.

[246] Mellers et al., "How Generalizable Is Good Judgement? A Multi-Task, Multi-Benchmark Study," 379.

[247] "Predicting the Future: A Lecture by Philip Tetlock," YouTube video, 1:15:16, posted by American Enterprise Institute, October 19, 2015, https://www.youtube.com/watch?v=xBXDTQdmNyw.

[248] Tetlock et al., "Forecasting Tournaments," 290–295.

[249] Tetlock et al., "Forecasting Tournaments," 291.

[250] David Ignatius, "David Ignatius: More Chatter than Needed," *Washington Post*, sec. Opinions, November 1, 2013, https://www.washingtonpost.com/opinions/david-ignatius-more-chatter-than-needed/2013/11/01/1194a984-425a-11e3-a624-41d661b0bb78_story.html.

III. A THOUGHT EXPERIMENT

A. IMAGINING A SUPERFORECASTED SELECTION MODEL

The goal of the following section is to construct a conceptual model representing how a superforecasted leadership selection process might function in practice, and then install it in the most optimal environment imaginable as a means of testing the hypothesis that it might work. A university classroom will serve as the backdrop for this exercise because it is instantly recognizable for most readers, and because classrooms are fairly static environments with readily measurable outputs. Also, university professors are reasonable analogies for leaders. Professors stimulate learning and engagement, provide vision, articulate goals, organize collective effort, drive performance, and reward achievement. Also, universities are currently facing some significant challenges and unintended consequences related to performance measurement that make them attractive candidates for inclusion in this thesis.

1. An Admittedly Indulgent Primer on a Problem in Academia

There is a longstanding and increasingly pervasive tension that exists within universities fomented upon the competing goals of teaching and research.[251] Universities are generally charged with producing knowledge, so cultivating high quality, high impact scholarly research is a critical function. However, even the most intensively research-focused institutions are expected to produce high quality students equipped with the requisite knowledge and critical thinking skills sufficient to further their respective fields and make a positive social contribution. The correlations observed between those different functions are not particularly strong.[252]

---

[251] John Hattie and H. W. Marsh, "The Relationship between Research and Teaching: A Meta-Analysis," *Review of Educational Research* 66, no. 4 (1996): 507–42, https://doi.org/10.2307/1170652; Simon Cadez, Vlado Dimovski, and Maja Zaman Groff, "Research, Teaching and Performance Evaluation in Academia: The Salience of Quality," *Studies in Higher Education* 42, no. 8 (August 3, 2017): 1455–73, https://doi.org/10.1080/03075079.2015.1104659.

[252] Herbert W. Marsh and John Hattie, "The Relation between Research Productivity and Teaching Effectiveness: Complementary, Antagonistic, or Independent Constructs?," *The Journal of Higher Education* 73, no. 5 (2002): 603–41, https://doi.org/10.1353/jhe.2002.0047.

In recent years, a crescendoing call from public and private organizations for universities to refocus their priorities on a teaching mission has been raised, driven in part by the increased demand for science, technology, engineering, and mathematics (STEM) professionals.[253] Nevertheless, and despite most universities outward assertions of being student-centered institutions of higher learning, faculty continue to be hired, evaluated, rewarded, and promoted on the basis of research.[254]

In a recent article for *University World News*, Philip Altbach and Hans de Wit suggest that the problem is magnified by national and global university ranking systems predicated solely upon perceptions of research prestige, leading to an isomorphic *me too* neurosis in which "most academic institutions want to resemble universities at the top of the academic pecking order."[255] This, the authors opine, has led to "a growing trend in doctoral education…to dispense with the traditional PhD dissertation and replace it with the requirement for doctoral students to publish several articles based on their research in academic journals."[256] By discarding dissertations for publications, universities hope to pump their rankings, thereby ostensibly pumping both research funding, and enrollment from students wishing to be associated with a perceptually prestigious institution. The result is a glut of submissions for publication from academics, leading to "a crisis in academic publishing—too much pressure on top journals, too many books of marginal quality, the rise of predatory journals and publishers that publish low or marginal quality research."[257] Academic journals have found themselves "buried in a barrage of papers," which has in recent months led one respected publication (citing a two-year backlog) to

---

[253] Stephen E. Bradforth et al., "University Learning: Improve Undergraduate Science Education," *Nature News* 523, no. 7560 (July 16, 2015): 282, https://doi.org/10.1038/523282a.

[254] Cadez, Dimovski, and Zaman Groff, "Research, Teaching and Performance Evaluation in Academia," 1455; Bradforth et al., "University Learning," 282.

[255] Philip Altbach and Hans de Wit, "Too Much Academic Research Is Being Published," *University World News*, September 7, 2018, http://www.universityworldnews.com/article.php?story=20180905095203579; "Tree for Two," YouTube video, 0:44, posted by Friz Freleng, Warner Brothers, 1952, https://www.youtube.com/watch?v=UVNHcob3oJg.

[256] Altbach and de Wit, "Too Much Academic Research Is Being Published."

[257] Altbach and de Wit.

suspend acceptance of new submissions, a decision decried by some academia insiders as "unheard of" and "simply stunning."[258]

This *overwhelming* effect has rippled to peer review, considered by many to be an unassailable hallmark of academic rigor, integrity, and veracity.[259] Consider, for example, one scorekeeping initiative undertaken of late by a team of 270 scientists from the Center for Open Science (COS) to validate "landmark" findings deemed statistically significant by 100 peer reviewed psychology studies, and published in three respected science journals.[260] Nearly 70% of those findings did not replicate, resulting in an "acrimonious" (albeit not surprising) public outcry of indignation from the implicated parties.[261] Undeterred, COS responded by evaluating the findings in several leading economics publications (40% non-replicable), followed by a consideration of multiple "exciting, innovative and important" studies published in two different journals of high regard (*Nature* and *Science)*. The latter endeavor revealed that over one third of those findings were not reproducible, and suffered from "false positives and inflated effect sizes" driven by "publication or reporting biases."[262]

---

[258] Tricia Serio, "Opinion: Repairing Peer Review," *The Scientist Magazine*®, November 18, 2016, https://www.the-scientist.com/opinion/opinion-repairing-peer-review-32512; Alexander C. Kafka, "Why Does Publishing Higher-Ed Research Take So Long?," *The Chronicle of Higher Education*, August 16, 2018, https://www.chronicle.com/article/Why-Does-Publishing-Higher-Ed/244291.

[259] Serio, "Opinion."

[260] Christian Jarrett, "Estimating the Reproducibility of Psychological Science," *The British Psychological Society Research Digest*, August 27, 2015, https://digest.bps.org.uk/2015/08/27/this-is-what-happened-when-psychologists-tried-to-replicate-100-previously-published-findings/; "About the Center for Open Science," Center for Open Science, accessed September 11, 2018, https://cos.io/.

[261] Jarrett, "Estimating the Reproducibility of Psychological Science"; Ed Yong, "A Failed Replication Draws a Scathing Personal Attack from a Psychology Professor," Discover Magazine, *Not Exactly Rocket Science* (blog), March 10, 2012, http://blogs.discovermagazine.com/notrocketscience/2012/03/10/failed-replication-bargh-psychology-study-doyen/; "We've Gotta Protect Our Phony Baloney Jobs!," YouTube video, 0:20, posted by Ed Morrissey, December 23, 2010, https://www.youtube.com/watch?v=uTmfwkl FM-M.

[262] Colin F. Camerer et al., "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015," *Nature Human Behaviour* 2, no. 9 (September 2018): 637–44, https://doi.org/10.1038/s41562-018-0399-z.

> It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.[263]

The source of those biases, according to Carnegie Mellon researchers Leslie John, George Loewenstein, and Drazen Prelec, may flow from a viciously erosive and self-sustaining cycle. Universities, incentivized by the prospect of improved prestige rankings, drive faculty members to publish in ever-increasing volume in academic publications, which are dis-incentivized by virtue of volume to publish studies with negative results.[264] This, in turn, incentivizes academics to engage in questionable research practices, described as "steroids of scientific competition, artificially enhancing performance," until the desired (and thereby publishable) positive result is achieved.[265]

A powerfully blunt critique of the contemporary university landscape was recently tweeted by terrorism powerhouse Bruce Hoffman, who bemoaned "the hyperventilating publish or perish culture in academe & erosion of teaching in stampede to publish. Especially lamentable is proliferation of PhD degrees eschewing traditional dissertations in favor of a series of articles/papers."[266] Dr. Hoffman's perspective is not benefitted by additional commentary from this writer.

For universities that do consider faculty teaching performance, at least in part, the overwhelmingly dominant method of assessment has been end of term student evaluations,

---

[263] Arthur Doyle, "A Scandal in Bohemia," in *The Adventures of Sherlock Holmes*, Project Gutenberg (Seattle, WA: Amazon Digital Services LLC, 2011): loc. 74 of 4318, Kindle, https://www.amazon.com/Adventures-Sherlock-Holmes-Arthur-Conan-ebook/dp/B06XPLKCSB/ref=tmm_kin_swatch_0?_encoding=UTF8&qid=&sr=.

[264] Leslie K. John, George Loewenstein, and Drazen Prelec, "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling," *Psychological Science* 23, no. 5 (May 2012): 524–32, https://doi.org/10.1177/0956797611430953.

[265] John, Loewenstein, and Prelec, 524; Brian Resnick and Julia Belluz, "A Top Cornell Food Researcher Has Had 13 Studies Retracted. That's a Lot," Vox, September 21, 2018, https://www.vox.com/science-and-health/2018/9/19/17879102/brian-wansink-cornell-food-brand-lab-retractions-jama.

[266] Bruce Hoffman, Twitter Post, September 9, 2018, 4:16 AM, https://twitter.com/hoffman_bruce/status/1038748067930009600.

which have in turn been resoundingly debunked as single-source assessment tools.[267] Why then, in the face of such damning evidence, would otherwise incomparably astute institutions continue to knowingly rely upon an ineffective metric to make profoundly consequential decisions about a critical infrastructure?[268] One erudite enclave of educators, including Dartmouth economist Douglas Staiger, politely suggest the reason may be driven by a heuristic of attribute availability, not unlike "the well-known story of a man looking for his keys under a street light—not because he dropped them nearby, but because that is where he can see."[269]

In response to the inefficacy of student evaluations as a single-source metric for professor proficiency, the RAND Corporation and others have called instead for an outcome based *value added model* (VAM), which seeks to track a teacher's contribution to undergraduate learning via longitudinal exam performance throughout the course of a given class.[270] This is a superficially attractive option, if for no other reason than it eliminates the subjectivity inherent in student opinion; however, much like end of course evaluations, it is not a method free from complication. John Ewing, then president of *Math for America*, enumerated a number of concerns related to VAM in a 2011 article published by the *American Mathematical Society*.[271] Ewing suggests that scores are influenced by multiple environmental factors beyond teaching proficiency, such as peer dynamics and

---

[267] Henry A. Hornstein, "Student Evaluations of Teaching Are an Inadequate Assessment Tool for Evaluating Faculty Performance," ed. Hau Fai Edmond Law, *Cogent Education* 4, no. 1 (March 20, 2017), https://doi.org/10.1080/2331186X.2017.1304016; Bob Uttl, Carmela A. White, and Daniela Wong Gonzalez, "Meta-Analysis of Faculty's Teaching Effectiveness: Student Evaluation of Teaching Ratings and Student Learning Are Not Related," *Studies in Educational Evaluation* 54 (September 2017): 22–42, https://doi.org/10.1016/j.stueduc.2016.08.007.

[268] "Thomas Mackin: Top Threats to America's Infrastructure," YouTube video, 14:59, posted by Center for Homeland Defense and Security Naval Postgraduate School, March 1, 2017, https://www.youtube.com/watch?v=7L3zlYavPE0.

[269] Jonah E. Rockoff et al., *Can You Recognize an Effective Teacher When You Recruit One?* (Cambridge, MA: National Bureau of Economic Research, 2011), 44, http://www.nber.org/papers/w14485.

[270] Roger Benjamin and Richard Hersh, "Measuring the Difference College Makes: The RAND/CAE Value Added Assessment Initiative," *Peer Review* 4, no. 2 (January 2, 2002), https://www.aacu.org/publications-research/periodicals/measuring-difference-college-makes-randcae-value-added-assessment.

[271] John Ewing, "Mathematical Intimidation: Driven by the Data," *Notices of the American Mathematical Society* 58, no. 5 (May 2011): 667–73, http://www.ams.org/notices/201105/rtx110500667p.pdf.

parental support, which cannot be captured by an exam.[272] Exams are only samples in a sense, much like polls, and therefore subject to the same snapshot statistical errors relating to the law of small numbers.[273] Exams are tests of a student's ability to learn facts and procedures; however, those tangible abilities are only a subset of learning objectives, which also include intangible attributes, such as "attitude, engagement, and the ability to learn further on one's own," which resist quantification.[274] The most challenging aspect to overcome, according to Ewing, is inflation, or the inherent incentive for professors to provide students with effective test-taking strategies, "teach to the test," or commit outright fraud, in the interest of boosting terminal performance ratings to advance career aspirations.[275]

Compelling evidence suggests that the incentive aspect of Ewing's objections should not be capriciously discounted. That finding was one of several from a study conducted at the U.S. Air Force Academy (USAFA) by Scott Carrell and James West.[276] The intent of the study was to consider VAM and student evaluations collectively as measures of professor performance within a given class, with the added consideration of student *follow-on* performance in a subsequent class (so calculus 101 in light of calculus 201).[277] This obvious yet previously unconsidered dimension shined new light on a dim space. The USAFA undergraduate student body is a well-suited testbed for objectively considering performance questions, because all students are randomly assigned to professors for 30 mandatory core courses covering mathematics, social sciences, engineering, basic sciences, and humanities, each utilizing a common syllabus.[278] The tracked follow-on courses for each of the core disciplines are likewise mandatory and

[272] Ewing, 668.

[273] Ewing, 668.

[274] Ewing, 668.

[275] Ewing, 668.

[276] Scott E. Carrell and James E. West, "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," *Journal of Political Economy* 118, no. 3 (2010): 409–432.

[277] Carrell and West, 409.

[278] Carrell and West, 411, 414.

randomized, thereby eliminating errors emanating from self-selection and attrition bias.[279] Another useful USAFA feature is that the exams are graded by a pool of professors within a given discipline, which allowed researchers to "rule out the possibility that professors have varying grading standards for equal student performance"[280] It also ensured that "bleeding heart professors had no discretion to boost grades or to keep their students from failing their courses," which thereby produced study results "driven by the manner in which the course is taught by each professor."[281] The data set used a diverse group of 10,534 undergraduate students at the USAFA over a seven-year period (2000–2007), and the output was derived from data extracted from 2,820 separate classes taught by 421 unique professors.[282] The performance of professors teaching introductory courses was measured by student scores in contemporaneous core classes, student end of course evaluations, and subsequent student scores in tracked follow-on courses.

Not surprisingly, Carrell and West observed that there were significant and substantial differences in student scores in both contemporaneous and follow-on classes.[283] Some students performed better than others. There was a correlation between student performance in a contemporaneous class and the manner in which a student rated the quality of the professor who taught that course.[284] Students with higher grades awarded professors with higher performance evaluations. Thus, professors with higher contemporaneous class averages are identified by students as being better teachers than those with lower class averages. Viewed in a bubble, this result appears to support the argument that student evaluations are a defensible means of measuring the quality of a teacher. This is the point at which Carrell and West shine, because the study further revealed, "student evaluations are positively correlated with contemporaneous professor

[279] Carrell and West, 412.

[280] Carrell and West, 419.

[281] Carrell and West, 419.

[282] Carrell and West, 414–15.

[283] Carrell and West, 412.

[284] Carrell and West, 412.

value-added and negatively correlated with follow-on student achievement."[285] The inverse was also true. Professors of contemporaneous courses who teach to the test produce students with higher scores who in turn reward professors with higher performance evaluations. The facet that is not captured by either contemporaneous metric is *deep learning*, or the professor's contribution to a student's understanding of a topic or discipline beyond that which is required to achieve an acceptable score on a specific exam. Deep learning is rigorous, and requires greater effort from both students and professors, but the payoff is improved understanding resulting in better performance in follow-on courses. In considering all of the available data, the study found that overall, individual professors "significantly affect student achievement in both the contemporaneous course being taught and the follow-on related curriculum."[286]

In summary, teachers matter, and some teachers are better than others. Status quo evaluations seeking to measure the quality of a professor's performance in the vacuum space of a single metric are specious and ill-advised. Incentives are important aspects of teacher performance evaluation systems and demand careful consideration. Predicting how a candidate for a teaching position observed today will perform tomorrow is a complex multivariate task in which cause and effect are not tightly bound. It is unlikely that any expert would possess the requisite knowledge to consistently make accurate predictions. It is considerably more likely that small portions of useful information are broadly dispersed, such that one person might have one scrap and a different person may hold another. A professor selection system that aggregates and averages the beliefs of a large, diverse group of forecasters should outperform the status quo. A selection system that tracks and reports the correlations between anticipated outcomes and actual outcomes will create a closed data feedback loop, which should produce gradually improved outcomes over time.

2. An Educated but Nonetheless Imperfect Best Guess

Imagine that a large American research college dubbed Jasper State University (home of the Scarlett Knights) has initiated an audacious, but not wholly unprecedented,

[285] Carrell and West, 412.

[286] Carrell and West, 429.

plan to create a new class of full-time professors dedicated to teaching core curriculum classes to the undergraduate student body.[287] If hired, this new faculty segment's sole task will be to produce the highest performing students possible, and future promotions will be dependent upon demonstrations of "excellence and leadership in education."[288]

JSU's first step is to recruit a large, diverse group of forecasters from its current, full-time employees in any rank or position; department chairs, HVAC technicians, physical therapists, or librarians. If warranted, JSU might broaden its recruitment pool to include alumni and retirees. The only prerequisite is a bachelor's degree from an accredited institution to reasonably ensure basic literacy. Such an initiative will require the digital distribution of an open solicitation which includes an overview of the initiative, as well as a compelling value proposition directly from the Chancellor emphasizing JSU's commitment to a radical, cutting-edge strategy designed to best prepare today's students for tomorrow's challenges. It should stress the opportunity for forecasters to test their cognitive mettle against the university's brightest minds in a meritocratic competition for meaningful recognition and greater civic good. Those who volunteer should do so with the understanding and expectation that they will receive ongoing training, support, and feedback while working collaboratively in teams of likeminded high-performers to better themselves, their school, and their community.

Once recruited, the forecasters will initially be randomly assigned to 10-person teams and loosely networked via a secure online Moodle platform designed to facilitate and memorialize forecaster interactions.[289] Once the teams are established, forecasters will complete an online block of instruction designed by JSU. Modeled after GJP's CHAMPS

---

[287] Mary Huber and Pat Hutchings, "New Teaching Positions up the Ante on Pedagogical Knowledge and Skill," *Bay View Alliance* (blog), February 26, 2018, http://bayviewalliance.org/new-teaching-positions-ante-pedagogical-knowledge-skill/.

[288] Leonard Cassuto, "A Tenure Track for Teachers?," *Chronicle of Higher Education*, May 7, 2017, https://www.chronicle.com/article/A-Tenure-Track-for-Teachers-/240015. It should be noted that tenure track versus non-tenure track positions in academia is a contentious topic that this paper does not seek to explore. Bradforth et al., "University Learning," 284.

[289] "Moodle Higher Education," Moodle, accessed September 18, 2018, https://moodle.com/higher-education/.

KNOW series, this course will serve as a best-practice guide for university forecasters, with graphically supported interactive segments relating to:

- Probabilistic reasoning
- Bayesian belief updating and Fermi estimations
- Common bias and heuristics errors and effective mitigation tactics
- Strategies to maximize the benefits of teams and overcome groupthink
- Strategies to test beliefs via Red Teaming and pre- and post-mortem analysis

Each segment will conclude with a knowledge check to allow forecasters to evaluate their understanding of the materials. The proceeding portion of the training will detail the means by which forecaster performance is to be measured, tracked, and reported, in addition to the method for measuring professor performance (discussed below). This includes an explanation of forecaster outcome accountability, as well as an overview of Brier scoring. The final segment will contain instructions to access a living list of shared reference links within the platform which forecasters can use and build upon in perpetuity to improve their understanding of a relevant topic area. The reference page will incorporate a feedback component, allowing forecasters to rate and comment upon the utility of a given resource, and might initially include links to recommended JDM research, GJP supeforecaster interviews, guidelines published by the Society for Industrial and Organizational Psychology (SIOP), or publications from the Yale Center for Teaching and Learning.[290]

[290] David Pinsen, "Interview with a Superforecaster," Seeking Alpha, February 10, 2016, https://seekingalpha.com/article/3882906-interview-superforecaster; "Superforecaster Full Video," YouTube video, 1:01:32, posted by KnowledgeAtWharton, February 26, 2016, https://www.youtube.com/watch?v=6POQjSjIXWk; "Welcome to SIOP," Society for Industrial and Organizational Psychology, Inc., accessed September 18, 2018, https://www.siop.org/; "Center for Teaching and Learning," Yale, accessed September 15, 2018, https://ctl.yale.edu/.

3. Data Input via Recruitment and Vetting

Once forecaster teams are established and trained, JSU then begins the process of recruiting applicants for teaching positions by distributing announcements for each of the core disciplines via status quo channels reflecting the new pedagogical focus. Applicants must possess a master's degree (MA) or Ph.D. from an accredited institution and may request consideration by submitting traditional documents to an online recruitment system, including a cover letter, curriculum vitae (CV), and any papers, published articles, other materials deemed by the applicant to be reflective of teaching excellence. In lieu of a research statement, applicants should submit a teaching statement which details teaching achievements, future goals, and concrete proposals for improving the quality of undergraduate education. Finally, all applicants must consent to a biometric criminal history query encompassing the United States and any foreign countries of residence.

Upon receipt of a complete electronic application packet, JSU human resources (HR) staff will perform an initial sift to verify that that the candidate meets minimum eligibility requirements. Next, an electronic copy of the packet is made and both files are assigned a unique digital tracking number. The original file (identity file) is encrypted and secured such that it cannot be altered. The duplicate application packet then undergoes an anonymization process, whereby any data which might identify the applicant's identity, gender, age, race, orientation, etc. is redacted. The anonymized file (forecaster file) is also encrypted and secured to prevent modification.

4. Narrowing the Candidate Pool

It is unreasonable for JSU, or any other large organization, to shoulder the time and expense of extensively vetting every eligible candidate for every open position, which is why all organizations have some mechanism (valid or not) for culling the best qualified from the merely qualified. It is at this point that university search committees, typically comprised of senior faculty and administrators, would begin to subjectively sort through CV piles and order-rank perceived potential; an unreliable process predicated upon storytelling and expert intuition. For any organization that requires a valid, legally defensible, single-metric mechanism to narrow a candidate pool, the literature

overwhelmingly supports general cognitive ability (GCA).[291] To date, following numerous independent studies, GCA is the single most reliable predictor of future workplace performance potential for any job type.[292] There are several standardized commercial instruments designed to measure GCA; however, JSU will use the Miller Analogies Test (MAT).[293] The MAT is broadly relied upon to screen candidates for admission to graduate schools and MENSA.[294] Further, a meta-analytic study conducted by Nathan R. Kuncel, Sarah A. Hezlett, and Deniz S. Ones found the MAT is a valid predictor for "academic and vocational criteria, as well as evaluations of career potential and creativity."[295] The MAT can accommodate persons with disabilities, and because the results are maintained for five years, many candidates will already possess a valid score at the time of application. For those that do not, the test is available throughout the world and requires no more than one hour to complete. Candidates with MAT scores in the 90% percentile or above may progress to a structured interview.

---

[291] Nicholas H. Morris, "A Review of Court Cases Involving Cognitive Ability Testing and Employment Practices: 1992–2015," paper 1575 (master's thesis, Western Kentucky University, 2016), 46, https://digitalcommons.wku.edu/cgi/viewcontent.cgi?article=2579&context=theses; Frank L. Schmidt, "The Role of General Cognitive Ability and Job Performance: Why There Cannot Be a Debate," *Human Performance* 15, no. 1 (2002): 187–210, https://www.researchgate.net/publication/240237107_The_Role_of_General_Cognitive_Ability_and_Job_Performance_Why_There_Cannot_Be_a_Debate.

[292] Schmidt and Hunter, "General Mental Ability in the World of Work."

[293] "Miller Analogies Test (MAT)," Pearson, accessed September 18, 2018, https://www.pearsonassessments.com/postsecondaryeducation/graduate_admissions/mat.html.

[294] "What Is Mensa?" Mensa International, accessed September 18, 2018, https://www.mensa.org/.

[295] Nathan R. Kuncel, Sarah A. Hezlett, and Deniz S. Ones, "Academic Performance, Career Potential, Creativity, and Job Performance: Can One Construct Predict Them All?," *Journal of Personality and Social Psychology* 86, no. 1 (2004): 148, https://doi.org/10.1037/0022-3514.86.1.148.

5. Structured Interviews (Version 2.0)

> One absolutely cannot tell, by watching, the difference between a .300 hitter and a .275 hitter. The difference is one hit every two weeks.[296]
>
> —Author Michael Lewis

In his 2003 book, *Moneyball*, Michael Lewis tells the tale of Billy Beane, the general manager of the Oakland Athletics, whose talent spotting acumen was fueled by one rule, never meet the talent. Surrounded by a sea of competing teams fervently devoted to the belief that their scouting instincts and expertise were the only keys to success, Beane quietly built multiple high performance teams on a shoestring budget, by recruiting players that others passed over because they did not conform to their concept of what a pro player should look like. Rather than watching a prospect perform and intuitively rating the performance, Beane built a predictive model, populated it with the player's performance statistics, and followed the data to a decision.[297] In a similar vein, symphonies around the world have dramatically improved audition outcomes by simply introducing an opaque screen into the equation so raters cannot observe the rated.[298] Years ago, Paul Meehl wagered that in a contest between doctors who can see patients, and computers who cannot, the smart money was on the machine; resulting in a longshot payout that continues to bear fruit.[299] Across multiple domains, the song remains the same. Most experts believe they can spot the goods. Most experts are wrong.

In light of the above considerations, JSU candidates who successfully hurdle the MAT will proceed to a structured interview stage conducted in strict conformation with

---

[296] Michael Lewis, *Moneyball: The Art of Winning an Unfair Game*, 1st. pbk. ed. (New York: Norton, 2004).

[297] J. Scott Armstrong, "Predicting Job Performance: The Moneyball Factor," *Foresight: The International Journal of Applied Forecasting*, 31–34, 2012, https://faculty.wharton.upenn.edu/wp-content/uploads/2012/05/Moneyball-Foresight_1.pdf.

[298] Claudia Goldin and Cecilia Rouse, "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians," *The American Economic Review* 90, no. 4 (2000): 715–41, http://www.jstor.org/stable/117305.

[299] Grove et al., "Clinical versus Mechanical Prediction."

industry standard guidelines established by SIOP, with one significant variation.[300] Those tasked with evaluating a candidate's interview performance will never see, hear, or speak with the candidate. Traditionally, interview panelists have two tasks, ask candidates job-related questions, and rate the quality of the responses received. JSU's method will use trained panelists drawn from appropriate university staff, who are exclusively responsible for eliciting the best possible information from every interview. In order to provide additional data points and promote interviewer buy-in, panelists are encouraged to notate their opinions about each candidate; however, forecasters may ultimately incorporate or disregard those beliefs at their discretion. All interviews are digitally recorded and transcribed into a text document (along with panelist notes), which is then anonymized, encrypted, and digitally merged with the candidate's forecaster file.

6. Data Processing and Output via Selection Forecasting

At this point in the process, every candidate has an anonymized file identified by a unique tracking number, which is digitally linked to an identity file. Anonymized files are populated with a CV and cover letter, a teaching statement with supporting documents, as well as an interview transcript with associated panelist notes. The anonymized files are then chunked into three sub-files (CV, teaching statement, and interview), which are randomly distributed to forecasters via the Moodle website. Working collectively with their assigned teams in a virtual environment, each forecaster is tasked with evaluating the contents of every sub-file and answering one question. If the candidate associated with this portion were selected for a position, would he or she receive a performance rating of at least 80% following two consecutive teaching semesters? Every forecast requires a binary yes/no prediction, accompanied by a confidence interval, so a forecaster evaluating a candidate's CV might predict YES 64% or NO 33%. Throughout this stage, forecasters are strongly encouraged to memorialize their thought processes and beliefs within the Moodle platform. Doing so will establish a cognitive crumb trail to help evaluate the quality of their judgments (and avoid *hindsight bias*) once feedback is available. In evaluating the

[300] "Effective Interviews," Society for Industrial and Organizational Psychology, Inc., accessed September 18, 2018, http://www.siop.org/workplace/employment%20testing/interviews.aspx.

interview transcripts, forecasters are likewise encouraged to notate their beliefs about the quality of the questions and the panelists. What information was or was not helpful? Are there questions that were not asked that should be incorporated into subsequent interviews? Are there opportunities to improve? To better visualize the flow, imagine five candidates for a JSU professor position. Each candidate has an anonymized file chunked into three sub-files, for a total of 15 sub-files. Those sub-files are randomly shuffled and digitally distributed to 100 forecasters. Every forecaster will evaluate the contents of each and submit 15 predictions, one for each sub-file, resulting in a combined total of 1,500 predictions to fill one position. The predictions for every sub-file are aggregated and averaged, producing three separate predictions for each candidate, which are aggregated and averaged again to derive a final estimate of future performance representing the collective wisdom of the crowd. The candidate with the highest performance prediction is awarded the position, begins the onboarding process, and proceeds to the lectern. This process repeats as necessary throughout the first year to fill every professor position announced by JSU for its core curriculum classes.

7. Measurement via Performance Management

The quality of a process can be determined by measuring the relationship between an anticipated outcome and an actual outcome. The resulting correlation represents critical feedback data that facilitates calibration in order to gradually produce better subsequent outcomes over time. This is true of engineering, chemistry, computer science, and personnel selection. For JSU, measuring and thereby improving the performance of its forecasters requires judging the performance of new professors by measuring outputs from the classrooms they were tasked to lead.

End of term student assessments are ineffective single-source metrics of professor performance because they are tremendously noisy instruments and the static overwhelms the signal, but valuable signal nonetheless resides therein. Assessments can capture the degree to which students perceive themselves to be engaged with the subject material and other less tangible classroom effects that impact learning, but are not necessarily reflected by raw test scores alone. Conversely, a value-add approach which only considers

longitudinal raw scores is equally undesirable as a solitary metric for largely the same reasons, but likewise carries valuable signals of professor performance. Both methodologies are also burdened by erosive incentivization issues which are ultimately harmful to students beyond the contemporaneous class, which must be confronted if the ultimate goal is longtime student learning and future societal contribution capacity. An admittedly appealing option would be to follow the surefooted path blazed by Carrell and West in their USAFA study by incorporating student follow-on performance; however, two years is too long and feedback delayed is feedback denied. How then to proceed? In their 2011 study, Rockoff et al. observed that "While no single metric we examine has the ability to reliably identify large differences in teacher effectiveness, we document these metrics can be used to create composite measures…which have statistically significant relationships with student achievement."[301] This finding parallels the foundational work of Clemen and others, thereby collectively lending credence to an old axiom, *Don't throw out the baby with the bathwater*.[302] JSU then will measure professor performance by aggregating and averaging the outputs produced by three measurements over the course of one academic year to form a composite reflection of anticipated outcome versus actual outcome.

The first measure is VAM, representing student scores for every core curriculum class taught by the new professor, and in a manner reflecting the professor's discretion, over the course of two consecutive semesters. The second measure is a 10-question end of term assessment designed to measure the extent to which a student's perception of class outcomes conforms with JSU expectations. Typically, such assessments require students to respond to each item using a 5-point Likert scale (strongly agree to strongly disagree) on the historical belief that people are only capable of discerning meaningful distinctions or degrees of ambiguity using coarse ordinal scales. That belief may not be universally accurate.

[301] Rockoff et al., *Can You Recognize an Effective Teacher When You Recruit One?*, 45.

[302] Clemen, "Combining Forecasts," 559.

One underappreciated finding of the GJP study reported by Jeffrey Friedman et al. was, "coarsening numeric probability assessments in a manner consistent with common qualitative expressions…consistently sacrifices predictive accuracy."[303] GJP researchers arrived at this conclusion by evaluating 888,328 probability assessments submitted by forecasters using a 1–100 confidence scale, and then artificially coarsening each prediction to conform with various status quo scales by rounding to the nearest interval bin.[304] Once the bins were populated, the Brier scores were recalculated. This exercise was repeated multiple times using three-, five-, and seven-point scales. Without exception, researchers found that coarsening "sacrifices meaningful information" and decreases accuracy.[305]

With this insight in hand, JSU will break with tradition and allow students to respond to each question using a 100-point scale as a means of imparting greater granularity, and thereby, in theory, capturing a greater portion of the available signal. At the end of an academic year, the results of the assessments will be aggregated and averaged first longitudinally, and then latitudinally, to produce a single cumulative result. A new professor teaching three core curriculum classes to 50 students per semester would produce 300 assessments comprised of responses to 10 questions, for a combined total of 3,000 data points. The mean value for each assessment item is calculated to produce 10 values, which are averaged again to produce a final result representing the collective perceptions of all students.

JSU will use a third and final measurement which is purpose-built to mitigate the incentive problem in the form of a blind exam. Studies have concluded that rating professors by either student evaluations or VAM creates the potential for teachers to inflate results by making students happy. This might include a range of manipulations, including teaching to a test, diminished workloads, eschewing instruction designed to promote deep learning, lenient grading, or possibly fraud. One way to check this problem might be to

---

[303] J. Friedman et al., "The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament," *International Studies Quarterly* (forthcoming) (2017): abstract, http://sites.dartmouth.edu/friedman/files/2017/06/Value-of-Precision-June-17.pdf.

[304] Friedman et al., 14–15.

[305] Friedman et al., 13–19, 32.

limit a professors' pedagogical discretion by mandating the use of a generic syllabus or standardized tests, but doing so limits a professors' academic freedom to teach and experiment in the classroom to the extent of her or his ability, and runs counter to the objectives served by acquiring high-performance professors in the first place. Accordingly, new JSU professors will be afforded absolute discretion over every aspect of student instruction save one, the final exam. At the end of each semester, every core curriculum class taught by a new professor will receive an exam which is collectively written and graded by follow-on professors, representing their estimation of the requisite knowledge and understanding required for success in the subsequent class. For new professors, this exam represents a blind process for which they have no knowledge or input.

8. Feedback

At the end of the first academic year, the three measurements (value-add, student evaluations, and the blind exam) are combined and averaged to produce a composite performance rating. This is an unquestionably imperfect rating process, but nonetheless superior to the single-metric solutions employed by universities today. Once the rating is obtained, Brier scores representing forecaster accuracy can be computed and fed back to the Moodle platform. Doing so sets several wheels in motion. It populates a leaderboard, which publicly ranks the performance of every forecaster and every team as a means of spurring rigor via recognition and competition. It also allows forecasters to evaluate the quality of their judgments and identify errors or opportunities for improvement. Rankings enable team rosters to be re-shuffled using a performance stratification scheme. Brier scores can be used to weight the value of a forecaster's subsequent prediction based upon the degree of accuracy observed in their prior performance. Hence, a forecaster who was 10% above the mean yesterday will have the value of a subsequent prediction increased by that same margin. All of the forecasters' comments and suggestions pertaining to the quality of the information produced by the interviews can be fed back to administrators as a means of identifying opportunities to improve the questions. Prior to the start of the next academic year, the new teams will complete another online block of training, and the process begins again.

9. Outcomes

The net result for JSU is a superforecasted selection process for new professors predicated upon the understanding that its underlying assumptions might be wrong, and must, therefore, be constantly tested and updated based upon new information. Transforming professors into distributed data-producing sensor arrays produces feedback, which can be used to correlate predicted outcomes with actual outcomes. This facilitates calibration via learning. Superforecasting enables the university's hiring system to identify mistakes in judgment and correct them in order to make better subsequent decisions. It is the creative application of the scientific method to personnel selection.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. PROBLEMS, BARRIERS, AND ISSUES FOR IMPLEMENTATION

> Those who believe that what you cannot quantify does not exist also believe that what you can quantify, does.[306]
>
> —Author Aaron Haspel

A. METRIC MISMANAGEMENT

The contemporary challenges and unintended consequences associated with how institutions of higher learning measure performance tell a cautionary tale of organizational dysfunction that will likely strike a resonant chord with many readers. Two valuable lessons can be drawn from the text. First, incomparably intelligent and well-intended organizations are not immune to profound judgmental flaws, and where issues of confidence are concerned, may actually be more susceptible to error if left unchecked. Second, numbers can be dangerous, even deadly, and the cascade effects of metric mismanagement can spread like a virus. DHS and many other organizations rely on the collective output from universities to guide and inform consequential decision-making; thus, a performance system that incentivizes volume over veracity raises some uncomfortable questions.

Universities, however, do not occupy the sole seat at the table of unintended consequences. In a recently published book entitled, *The Tyranny of Metrics*, historian Jerry Muller identifies a number of institutions whose pernicious use of performance metrics has resulted in suboptimal outcomes. In law enforcement, the collection and use of data has been invaluable as a means of resource allocation, trend mapping, and organizational learning which has been increasingly leveraged in recent years to great effect. However, statistics have also been political implements used to sway public perception or bolster reputational currency, producing a trickle-down effect that

[306] Jerry Z. Muller, *The Tyranny of Metrics* (Princeton, Oxford: Princeton University Press, 2018), epigraph quoting Aaron Haspel, https://www.amazon.com/Tyranny-Metrics-Jerry-Z-Muller/dp/0691174954.

incentivizes gamification in the form of underreporting, downgrading, or effort diversion.[307] Such manipulations formed the basis for the fictionalized story arc presented in the award-winning HBO series *The Wire*.[308] Muller suggests another example in the Vietnam era U.S. military, and the influence of Defense Secretary Robert McNamara, a career accountant, who established metrics of success predicated upon bomb drops and body counts.[309]

> All who drink of this treatment recover in a short time, except those whom it does not help, who all die. It is obvious, therefore, that it fails only in incurable cases.[310]

One of the most illustrative fields regarding the duality of metrics is medicine. Until rather recently (in historical terms), it was an industry driven by intuition, superstition, personality, and perception, such that "it was not unusual for a sick person to be better off if there were no physician available because letting an illness take its natural course was less dangerous than what a physician would inflict."[311] This state of affairs persisted until the 20th century, when the introduction of objective observations to correlate anticipated outcomes with actual outcomes via randomized experimentation dramatically accelerated advancement within the field. The flip side of that coin, however, can be observed in some modern medical institutions and the use of performance metrics represented by report cards or star-ratings as single-source forms of physician or facility accountability. According to Muller, whether well-intended or not, these measures have incentivized the gamification of the process, leading to (for example) case selection bias from risk-averse physicians, or hospitals leaving patients in ambulances to stall the clock in response to a performance metric measuring admittance wait times.[312]

---

[307] Muller, 125–29.

[308] "The Wire," IMDb. accessed October 7, 2018, http://www.imdb.com/title/tt0306414/.

[309] Muller, *The Tyranny of Metrics*, 131–35.

[310] Druin Burch, *Taking the Medicine: A Short History of Medicine's Beautiful Idea, and Our Difficulty Swallowing It* (London: Vintage Books, 2010), loc. 37 of 4951, Kindle, quoting Galen.

[311] Tetlock and Gardner, *Superforecasting*, 26.

[312] Muller, *The Tyranny of Metrics*, 5, 103–23.

Measurement is not a panacea. Many groups serve multiple purposes. It is impossible to measure every aspect of a group's performance, and measuring one or two facets may be an invitation to neglect other important group functions. The virtue of a closed-loop data feedback network like the one created by a superforecasting model is its capacity to enable a group to identify errors and learn from mistakes, which stimulates learning. Conversely, utilizing metrics as a means to punish, sway opinion, boost reputation, or support a narrative will likely incentivize bad behavior.

B. ORGANIZATIONAL RETICENCE

Superforecasting is a process predicated upon failure and an institutional acceptance of the fact that it is going to be wrong. Forecasting is, in part, dependent upon negative outcomes to stimulate learning, so the acceptance of a *fail forward* mindset prior to implementation is critical.[313] Eighty percent predictive accuracy in domains of complexity is an amazing mark to attain. However, 80% right is still 20% wrong, meaning that a highly skilled team of superforecasters is going to miss the mark one time in five. For organizations accountable for the effective stewardship of public funds, the notion of intentionally adopting a process that is going to result in failures may be a difficult barrier to breach.

Adopting a superforecasting model for leadership selection would also require existing leaders to completely abdicate control of a powerful hallmark of authority, the ability to promote (and thereby reward) subordinates, and transfer that capacity to a distributed network of unknowns. For leaders whose identity is inextricably intertwined with the social perception of themselves as horse borne battlefield commanders decisively waging war, rather than McChrystal's model of a holistic gardener organically cultivating the organizational capacity for greatness to spring and prosper, relinquishing the reins will be a challenging proposition to accept, in spite of its value.[314]

---

[313] Richard Farson and Ralph Keyes, "The Failure-Tolerant Leader," *Harvard Business Review*, August 1, 2002, https://hbr.org/2002/08/the-failure-tolerant-leader.

[314] McChrystal et al., *Team of Teams*, chap. 11.

C. THE ROAD NOT TRAVELLED

Tetlock's forecasters predicted future outcomes of highly complex geopolitical events. Feedback in the form of a Brier score represented the variation between what a forecaster believed would happen and what actually happened. Thus, forecasters worked in a truly dichotomous environment in which the only possible outcomes were yes or no, and judgments could be accurately assessed as a means of confirming mental models or identifying flaws. A superforecasted leadership selection system could only assess outcomes directly associated with the proceeds of the process. What about the candidates who were not selected for promotion?

Such counterfactual questions are important points for consideration. If a candidate ultimately performed well in conformation with a forecaster's prediction, then a Brier score reflecting good judgment is awarded and a forecaster can update beliefs accordingly. However, consider the possibility that an overlooked candidate for promotion would have performed better, had she or he been chosen. If that were to be true, then the forecaster actually made a mistake and exercised poor judgment. Without a means of accurately assessing all possible outcomes, it is impossible to know. This gap is not completely dispositive, because the extent to which a prediction correlates to an outcome remains an opportunity for learning; it simply means that learning potential is mediated by a network's capacity to observe all potential outcomes.

D. QUALITATIVE TASKS RESIST QUANTITATIVE MEASUREMENT

Some group functions and outputs are more measureable than others. Additionally, the work product of some groups is wholly qualitative. For DHS, meaningfully measuring the performance of U.S. Air Marshals tasked with securing America's airways, or members of U.S. Secret Service protective details, would be a futile exercise. Conversely, other groups with more process-oriented missions, such as instructor cadre at the Federal Law Enforcement Training Center, or security screeners within the Transportation Security Administration, produce observable outcomes which more readily lend themselves to measurement.

E. A STATISTICAL CONUNDRUM

The final barrier to implementing a superforecasting model for leadership selection is also the most intractable. This thesis suggests that group performance should be a proxy for leadership performance. However, even if the problems associated with qualitative measurement could be reasonably surmounted, most groups are not large enough to produce statistically significant results. Many leaders in DHS (and virtually every other organization) only supervise a handful of employees. Small samples produce extreme results and are not suitable for use as a basis for learning. Increasing the ratio of employees to leaders to conform to the laws of statistics would produce teams so large that effective leadership would be an impossible task. Without accurate feedback, forecasters could never effectively evaluate the quality of their predictions, which is a requisite feature of the superforecasting process.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSION

A. SUMMARY

No plan survives first contact. That sentiment was originally espoused by Prussian commander Helmuth von Moltke to describe the military's essential acceptance of the fact that no strategy will ever align with the inherent complexities of war. Moltke's adage could also aptly account for the circuitous trajectory of a thesis propelled by a theoretical exercise in belief updating.

DHS has a leadership problem. The IC has an analysis problem. Examining root causes for both conditions at a network level revealed surprisingly congruent fundamental design flaws relating to measurement, structure, and decision-making. The GJP started with a question. How can an organization make the best possible predictions of complex future events? That question evolved into a loosely held belief that a large, diverse group of crowdsourced forecasters would be more accurate than any other predictive methodology. Tetlock used the IARPA tournament to test that belief and refine it over time by following the outputs from a series of controlled experiments. The final product was superforecasting, which harnesses the capacity for algorithmically enhanced groups to evolve through diversity, training, performance stratification, feedback, recognition, and accountability. Tetlock's recipe was not perfect, but it was far better than the IC's status quo.

The causal overlaps for the flaws observed in geopolitical analysis and leadership selection suggested a novel possibility. DHS could use a superforecasting methodology to improve its leadership selection decisions in the same manner that GJP used superforecasting to make better geopolitical predictions. That assumption launched an exploratory thought experiment that synthesized the relevant literatures into a technical manual of sorts, to deconstruct Tetlock's machine and understand the mechanisms of operation. The following chapter modified the motor for personnel selection and installed it in the most advantageous environment imaginable to gauge its odds of success. Finally,

a Red Team of wrench-bearing monkeys was loosed to probe for weakness and foul the gears. They proved to be annoyingly shrewd simians.

Would superforecasting work part and parcel for DHS leadership selection systems? No. At least, not in a way that would make it generalizable for every workgroup. The problems associated with qualitative measurement, small numbers, and counterfactuals are beyond this writer's capacity to overcome.

The question, however, was not asked in vain. The process of exploring that loosely held belief to an unexpected result produced valuable insights that still make a positive contribution to the problem space. Tetlock's motor might not be a perfect theoretical fit for this specific application, but now that the mechanism is understood, component pieces can be disassembled and put to work for DHS.

This thesis identified systemic flaws in the DHS leadership selection process, and offers several concrete recommendations that can be implemented to produce better organizational outcomes. It also lays a crumb trail for others to follow, and potentially build upon. That is an acceptable result.

Leaders matter, and some leaders are better than others. Were that not true, organizations could forego the time and expense of a selection process and choose leaders by flipping coins or drawing straws. Leaders are born *and* made. Focusing on development at the expense of selection is an indication of poor judgment. Even marginal improvements to the way organizations choose leaders can pay a substantial dividend. Leadership selection is a complex domain in which cause and effect are not tightly bound. Were that not true, status quo leadership selection systems would produce optimal outcomes. Experts thrive in complicated domains and perform poorly in domains of complexity. Misapplying complicated solutions to complex problems will produce poor outcomes. Leadership selection is a prediction, an educated but nonetheless imperfect belief about how a candidate observed today will perform tomorrow. Beliefs are theories to be tested, not treasures to be cherished, and should be constantly updated based upon the best available data. Closed loop data systems perform better than open loop systems because correlating anticipated outcomes with actual outcomes generates feedback, which enables the learning

required to produce better subsequent decisions. Without feedback, decision makers are free to presume that all of their decisions are good decisions. This creates an environment not unlike a petri dish, where cognitive biases and bad heuristics can flourish. Good training is an excellent inoculant and good feedback is the best disinfectant. Searching for ways that leadership selection systems can learn and produce progressively better outcomes is a worthwhile endeavor.

What organizations think is *far* less important than how.

B. RECOMMENDATIONS

> If you don't know where you're going, you might end up somewhere interesting.[315]
>
> —Naval Postgraduate School professor Christopher Bellavita

1. Improving the DHS Leadership Selection Process

- Stop focusing on leadership development at the expense of selection. To date, DHS has essentially exerted tremendous effort chasing an errant group of horses around a pasture to which they should not have been admitted. DHS should stop running and build a better barn door. Some leadership candidates are better than others, so even modest improvements to the predictive accuracy of the selection process can pay significant organizational dividends.

- DHS must disabuse itself of the notion that experts can intuitively identify good leaders when they see them. They cannot. The moment that any expert successfully solves the cause/effect selection puzzle in a manner that is replicable and reliable, the multi-billion dollar leadership industry will cease to exist. Until then, leadership selection remains a complex task for which experts are ill-suited. A large, cognitively diverse group of

[315] Christopher Bellavita, "NPS-CHDS Capstone Lecture" (lecture, Naval Postgraduate School, Monterey, CA, October 5, 2018).

trained selection officials will produce better outcomes over time than a small homogenous group of experts.

- Every employee tasked by DHS with evaluating the future performance potential of a leadership candidate should receive an annual block of high quality training dedicated to judgment and decision making, probabilistic reasoning, cognitive bias, heuristics, effective teaming, belief updating, and groupthink avoidance.

- Because the performance of a leader is impossible to measure objectively, DHS should seek opportunities to use group performance as a proxy for leadership performance whenever possible.

- Make the system for evaluating leadership candidates a blind process by anonymizing application materials to remove any biographic indicators. There is no predictive value to be found in a candidate's name, gender, race, ethnicity, etc. Allowing evaluators to have access to that data is an invitation to unconscious cognitive errors. Prior to a final selection decision, officials should be judging the future performance potential of multiple barcodes rather than multiple people.

- Candidates for first-line leadership positions should be evaluated based upon their capacity to perform the job they are competing for, rather than the job they currently have. Past performance only predicts future performance within solitary domains. Contributing to an effective team and leading an effective team are different outcomes, requiring different proficiencies. Measuring technical proficiency is easy. Easy does not equal effective.

- Design performance management systems that produce meaningful data to potentiate learning and improvement. A feedback loop that identifies every employee as fully successful has no value.

- In order to improve the quality of the candidate pool for leadership positions, DHS should consider implementing a dual pipeline for employee promotions beyond the journeyman grade. Currently, highly skilled employees are incentivized to apply for leadership positions they may not want because the alternative is career stagnation. Similarly, selection officials are incentivized to choose such high performers because no alternative means of recognizing or rewarding them exists. Allowing employees the flexibility to choose between technical and leadership career pipelines would improve both candidate pools, and allow DHS to groom employees for the positions they want, rather than the positions to which they were relegated to apply.

- As suggested by the JSU model imagined above, DHS should radically alter the way it interviews candidates for leadership positions by adopting a Moneyball methodology. Officials tasked with evaluating candidates should never have an opportunity to see or hear any candidate. Structured interviews should be conducted by a panel of trained interviewers whose sole task is the elicitation of the best possible information from the candidate in a valid, reliable manner. Anonymized interview transcripts can then be digitally distributed to trained selection officials for blind evaluations. Tasking one group of people with conducting *and* evaluating interviews is a recipe for poor judgment.

- In addition to anonymizing all selection packets to remove biographic information, the packets for every candidate should also be chunked into logical portions and randomized prior to evaluation. Chunking will mitigate the halo effect, or the unconscious tendency for positive or negative impressions formed by a rater in one area to unjustly influence the rating of an unrelated area.

- Use multiple selection tools to evaluate the performance potential of leadership candidates. Combining multiple forecasts increases predictive

accuracy. Selection processes that only use one or two data sources (e.g., resume and interview) are inherently less accurate than those using multiple data sources.

- Incorporate standardized GCA testing into the leadership selection process. GCA is the most reliable predictor of future workplace performance for all job types, and should be used more broadly by DHS to narrow candidate pools.

2. Improving Organizational Judgment and Decision Making

- Design a formal curriculum for the DHS leadership cadre dedicated to improving JDM. Research demonstrates that even modest periodic training investments dedicated to probabilistic reasoning, bias mitigation, belief updating, active open mindedness, post-mortem analysis, and groupthink avoidance can pay significant returns. JDM training should be included in the onboarding curriculum for new leaders, incorporated into mandatory annual training requirements for existing leaders, and made digitally available for all employees via the DHS intranet. At a minimum, every DHS leader should be familiar with the findings of Tversky, Kahneman, and Tetlock.

- Task the DHS Chief Learning and Engagement Officer with the creation of a JDM intranet resource page populated with best examples of video lectures, articles, research, books, and other digital media related to the topic. Seeking partnership with a leading academic institution, such as the University of Pennsylvania's Wharton School to assist in the effort, would be a reasonable first step.

- Understand the difference between complex and complicated problem spaces as a means of discerning the optimal path to a solution. If the relationship between cause and effect is observable and replicable, then using a diverse group to discern a solution is inefficient. In such cases,

machines or experts would produce better outcomes. Conversely, if a problem space is unpredictable such that the relationships between cause and effect are not tightly bound, then diversity will outperform expertise.

- Be wary of labels. Too often, people are anointed with the title of expert simply because they have been working in a particular field for a long period of time. Experience and expertise are not the same thing. Poor long-term performance is evidence of ignorance, not aptitude.

- Understand the systemic reasons why groups of complex problem solvers succeed or fail. Groups can work collaboratively to produce judgments which are more accurate than the best estimate of any single member, but only under specific conditions. Diversity, training, feedback, individual accountability, recognition, and algorithms are critical components for effective crowdsourcing because of their collective capacity to filter static, amplify signal, and progressively improve over time. Conversely, groups of otherwise incomparably intelligent problem solvers can produce judgments which are worse than those of any single member if they are left untrained, unaccountable, homogenous, consensus driven, and without effective feedback. When randomly wrong becomes consistently wrong, aggregating and averaging does not filter out the static; it amplifies it, thereby obscuring signal fidelity. A group that cannot learn from its mistakes is going to repeat them.

- DHS should refine the way it conceptualizes diversity. If pressed, many who insist that diverse groups are better than homogenous groups probably could not explain why, because touting diversity is a socially acceptable thing to do. Diversity is not a polite abstraction; it is a silver bullet. Diversity is a tangible, game changing force that can radically improve the performance of a group in the same way that a fulcrum can radically improve the performance of a lever. In order to function optimally a group needs to be cognitively diverse, yet most organizations

strive for biographic diversity; a myopic perspective that hampers potential. Biographic diversity is a valid path to cognitive diversity, but there are others. When forming groups to work in complex problem spaces, DHS should strive for cognitive diversity.

- Task the DHS Office of Science and Technology with implementing a secure intranet platform modeled after GitHub to enable enterprise-wide collaboration and idea sharing. DHS has an ever-changing range of complex challenges to navigate. It also has a large, talented, diverse workforce, tailor-made for developing innovative solutions. What the department lacks is a dedicated space to apply its people to its problems by harnessing the power of crowdsourcing and social recognition. It should be a meritocratic environment where identities are transparent, titles are irrelevant, and ideas are elevated via peer-ratings and feedback based solely on the quality of the contribution. Any DHS employee could use the site as a living white board to issue challenges or offer solutions, which others could then critique, incorporate, modify, or ignore. Such an environment would also provide DHS with a valuable opportunity to identify unrecognized talent that could then be repurposed to better serve both the mission, and the employee.

C. SUGGESTIONS FOR FUTURE RESEARCH

1. Boosting Accuracy via Affinity Weighting

If the historical data produced by every weather forecasting model throughout the world were analyzed to determine the accuracy of each, it seems reasonable to imagine that over time, some models might be found to have developed affinities for certain types of meteorological phenomena. For example, one model might be consistently above average at predicting microbursts and another might be reliably better at guessing snowstorm accumulations. During the IARPA tournament, Tetlock only considered the overall historical performance of each forecaster as a means of weighting the value of their subsequent prediction. It is possible that Tetlock missed an opportunity to algorithmically

extract a modicum of additional accuracy from his forecasters. What if, over time, human forecasters exhibited affinities for specific topic areas, like North American commodity markets or European elections? Were that true, then the value of a prediction could be weighted by overall performance, *as well as* performance within a specific domain. It is a testable hypothesis.

2. Boosting Accuracy via Granularity

Consider a counterfactual question: What fate would have befallen Surowiecki's fame had Galton forced his fairgoers to guess the bulk of the beast in 50-pound intervals? One notable finding that emerged from GJP was that forecasters are capable of meaningfully distinguishing between finer degrees of uncertainty than previously believed. In fact, there was a strong correlation observed between the granularity of a forecaster and the accuracy of a forecaster. Further, Tetlock found that artificially coarsening forecasts after the fact by rounding to the nearest interval sacrificed signal and eroded accuracy. As a result, some components of the IC have considered abandoning status quo scales of uncertainty in favor of more granular alternatives. Given that most leadership selection processes use traditional 5-point Likert scales to rate various aspects of a candidate's performance potential, would allowing raters to use a 100-point scale produce better results?

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Aisner, James. “Andy Grove: A Biographer’s Tale.” Harvard Business School, Working Knowledge, November 9, 2006. https://hbswk.hbs.edu/item/andy-grove-a-biographers-tale.

Altbach, Philip, and Hans de Wit. “Too Much Academic Research Is Being Published.” *University World News*, September 7, 2018. http://www.universityworldnews.com/article.php?story=20180905095203579.

American Enterprise Institute. “Predicting the Future: A Lecture by Philip Tetlock.” YouTube video, 1:15:16. October 19, 2015. https://www.youtube.com/watch?v=xBXDTQdmNyw.

Anderson, Chris. “Closing the Loop: A Conversation with Chris Anderson.” Edge. Accessed October 22, 2017. https://www.edge.org/conversation/chris_anderson-closing-the-loop.

Armstrong, J. Scott. “Predicting Job Performance: The Moneyball Factor.” *Foresight: The International Journal of Applied Forecasting*, 2012. https://faculty.wharton.upenn.edu/wp-content/uploads/2012/05/Moneyball-Foresight_1.pdf.

———. “The Seer-Sucker Theory: The Value of Experts in Forecasting.” *Technology Review* 82, no. 7 (June 1980): 16–24. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=648763.

Ashton, Robert H. “Effects of Justification and a Mechanical Aid on Judgment Performance.” *Organizational Behavior & Human Decision Processes* 52, no. 2 (July 1992): 292–306.

Ball, Philip. “The Trouble with Scientists.” *Nautilus*, May 14, 2015.

Bar-Hillel, Maya. “The Base-Rate Fallacy in Probability Judgments.” *Acta Psychologica* 44, no. 3 (May 1, 1980): 211–33. https://doi.org/10.1016/0001-6918(80)90046-3.

Bass, Bernard M., and Ruth Bass. *The Bass Handbook of Leadership: Theory, Research, and Managerial Applications*. 4th ed. New York: Free Press, 2008.

BBC Earth Lab. “How Do Noise Cancelling Headphones Work?—James May’s Q&A (Ep 10)—Head Squeeze.” Accessed December 26, 2017. https://www.youtube.com/watch?v=VTx4JgYsW5s.

Benjamin, Roger, and Richard Hersh. “Measuring the Difference College Makes: The RAND/CAE Value Added Assessment Initiative.” *Peer Review* 4, no. 2 (January 2, 2002). https://www.aacu.org/publications-research/periodicals/measuring-difference-college-makes-randcae-value-added-assessment.

Bernicker, Brendan. “Gedankenexperiment|Thought Experiments.” Penn State University, February 4, 2016. https://sites.psu.edu/bernickerpassionblog/2016/02/04/gedankenexperiment/.

Berson, Yair, Orrie Dan, and Francis J. Yammarino. “Attachment Style and Individual Differences in Leadership Perceptions and Emergence.” *The Journal of Social Psychology* 146, no. 2 (April 2006): 165–82. https://doi.org/10.3200/SOCP.146.2.165-182.

Betts, Julian R., and Jamie L. Shkolnik. “The Effects of Ability Grouping on Student Achievement and Resource Allocation in Secondary Schools.” *Economics of Education Review* 19, no. 1 (February 2000): 1–15. https://doi.org/10.1016/S0272-7757(98)00044-2.

Bloomberg. “John F. McCreary: Executive Profile & Biography.” Accessed July 13, 2018. https://www.bloomberg.com/research/stocks/private/person.asp?personId=267776426&privcapId=733543&previousCapId=733543&previousTitle=Kforce%20Government%20Solutions,%20Inc.

Bose. “Bose Noise Cancelling Headphones.” Accessed July 13, 2018. https://www.bose.com/en_us/products/headphones/noise_cancelling_headphones.html.

Bradforth, Stephen E., Emily R. Miller, William R. Dichtel, Adam K. Leibovich, Andrew L. Feig, James D. Martin, Karen S. Bjorkman, Zachary D. Schultz, and Tobin L. Smith. “University Learning: Improve Undergraduate Science Education.” *Nature News* 523, no. 7560 (July 16, 2015): 282–284. https://doi.org/10.1038/523282a.

Brier, Glenn W. “Verification of Forecasts Expressed in Terms of Probability.” *Monthly Weather Review* 78, no. 1 (January 1, 1950): 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brockman, John, Russell Weinberger, and Nina Stegeman. “Edge Masterclass 2015: A Short Course in Superforecasting, Class I.” Edge, August 24, 2015. https://www.edge.org/conversation/philip_tetlock-edge-master-class-2015-a-short-course-in-superforecasting-class-i.

Budescu, David V., and Eva Chen. “Identifying Expertise and Using It to Extract the Wisdom of the Crowds.” *Management Science* 61, no. 2 (May 23, 2014): 1–37. http://pages.stern.nyu.edu/~eyoon/seminar/dbudescu/Paper.pdf.

Burch, Druin, *Taking the Medicine: A Short History of Medicine’s Beautiful Idea, and Our Difficulty Swallowing It*. London: Vintage Books, 2010. Kindle.

Busch, Noel. “Close-Up: Lord Keynes.” *Life*, September 17, 1945. https://books.google.com/books?id=t0kEAAAAMBAJ&q=%22a+cable%22#v=snippet&q=%22a%20cable%22&f=false.

Cadez, Simon, Vlado Dimovski, and Maja Zaman Groff. "Research, Teaching and Performance Evaluation in Academia: The Salience of Quality." *Studies in Higher Education* 42, no. 8 (August 3, 2017): 1455–73. https://doi.org/10.1080/03075079.2015.1104659.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, and Michael Kirchler et al. "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour* 2, no. 9 (September 2018): 637–44. https://doi.org/10.1038/s41562-018-0399-z.

Cannon-Bowers, Janis A., and Eduardo Salas, eds. *Making Decision Under Stress: Implications for Individual and Team Training*. 1st ed. Washington, DC: American Psychological Association, 1998.

Caplan, Bryan. "Have the Experts been Weighed, Measured, and Found Wanting?" *Critical Review*, November 2, 2007.

Carrell, Scott E., and James E. West. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *Journal of Political Economy* 118, no. 3 (2010): 409–432.

Cassuto, Leonard. "A Tenure Track for Teachers?." *Chronicle of Higher Education*, May 7, 2017. https://www.chronicle.com/article/A-Tenure-Track-for-Teachers-/240015.

Center for Homeland Defense and Security Naval Postgraduate School. "Thomas Mackin: Top Threats to America's Infrastructure." YouTube video, 14:59. March 1, 2017. https://www.youtube.com/watch?v=7L3zlYavPE0.

Center for Open Science. "About the Center for Open Science." Accessed September 11, 2018. https://cos.io/.

Chang, Welton, Eva Chen, Barbara Mellers, and Philip Tetlock. "Developing Expert Political Judgment: The Impact of Training and Practice on Judgmental Accuracy in Geopolitical Forecasting Tournaments." *Judgment and Decision Making* 11, no. 5 (September 2016): 509–26. http://journal.sjdm.org/16/16511/jdm16511.pdf.

Clemen, Robert T. "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting* 5, no. 4 (1989): 559–583.

Clemen, Robert T., and Robert L. Winkler. "Combining Economic Forecasts." *Journal of Business & Economic Statistics* 4, no. 1 (January 1986): 39–46. https://doi.org/10.2307/1391385.

Colburn, Tom. *A Review of the Department of Homeland Security's Missions and Performance*. Washington, DC: Committee on Homeland Security and Governmental Affairs, 2015. http://www.hsgac.senate.gov/download/?id=B92B8382-DBCE-403C-A08A-727F89C2BC9B.

Crandall, Beth, and Karen Getchell-Reiter. "Critical Decision Method: A Technique for Eliciting Concrete Assessment Indicators from the Intuition of NICU Nurses." *ANS. Advances in Nursing Science* 16, no. 1 (September 1993): 42–51.

Cummins, Eleanor. "There Was another Earthquake in Mexico. Is the World Ending?" *Slate*, September 19, 2017. http://www.slate.com/articles/health_and_science/science/2017/09/this_summer_has_been_an_unending_series_of_natural_disasters.html.

Dabbish, Laura, Colleen Stuart, Jason Tsay, and Jim Herbsleb. "Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository." In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work—CSCW '12*, ACM 2012 conference. 1277–1286. Seattle: WA: ACM Press, 2012. https://doi.org/10.1145/2145204.2145396.

Dalyell, Tam. "Westminster Scene." *New Scientist* 60, no. 871 (November 8, 1973): 423–426. https://books.google.com/books?id=i145R0bZXMYC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false.

de Groot, Adriaan D. *Thought and Choice in Chess*. Berlin: Walter de Gruyter GmbH & Co KG, 1978.

de Langhe, Bart, Stijn M.J. van Osselaer, and Berend Wierenga. "The Effects of Process and Outcome Accountability on Judgment Process and Performance." *Organizational Behavior and Human Decision Processes* 115, no. 2 (July 2011): 238–52. https://doi.org/10.1016/j.obhdp.2011.02.003.

Department of Homeland Security. "DHS Leadership Year." December 6, 2017. https://www.dhs.gov/dhs-leadership-year.

Derue, D. Scott, Jennifer D. Nahrgang, Ned Wellman, and Stephen E. Humphrey. "Trait and Behavioral Theories of Leadership: An Integration and Meta-Analytic Test of Their Relative Validity." *Personnel Psychology* 64, no. 1 (2011): 7–52. http://onlinelibrary.wiley.com/doi/10.1111/j.1744-6570.2010.01201.x/full.

Doyle, Arthur. "A Scandal in Bohemia." In *The Adventures of Sherlock Holmes*. Project Gutenberg. Seattle, WA: Amazon Digital Services LLC, 2011. Kindle. https://www.amazon.com/Adventures-Sherlock-Holmes-Arthur-Conan-ebook/dp/B06XPLKCSB/ref=tmm_kin_swatch_0?_encoding=UTF8&qid=&sr=.

———. “The Adventure of the Copper Beeches.” In *The Adventures of Sherlock Holmes*, 2017. Kindle. https://www.amazon.com/Adventures-Sherlock-Holmes-Arthur-Conan-ebook/dp/B06XPLKCSB/ref=tmm_kin_swatch_0?_encoding=UTF8&qid=&sr=.

Duke, Annie. *Thinking in Bets: Making Smarter Decisions When You Don’t Have All the Facts*. New York: Portfolio Penguin, 2018.

Duke Today. “Who Was Thomas Bayes?” Accessed August 13, 2018. https://today.duke.edu/2012/11/bayes.

Eagly, Alice H., Mona G. Makhijani, and Bruce G. Klonsky. “Gender and the Evaluation of Leaders: A Meta-Analysis.” *Psychological Bulletin* 111, no. 1 (1992): 3–22. https://doi.org/10.1037/0033-2909.111.1.3.

Edge. “Edge Master Class 2015—Philip Tetlock: A Short Course in Superforecasting.” Accessed October 11, 2017. https://www.edge.org/event/edge-master-class-2015-philip-tetlock-a-short-course-in-superforecasting.

El Santo. “Kirk vs. Picard: Who’s the Best Starfleet Captain?.” *Rooktopia* (blog), May 16, 2013. https://rooktopia.wordpress.com/2013/05/16/kirk-vs-picard-whos-the-best-starfleet-captain/.

Epple, Dennis, and Richard E. Romano. “Peer Effects in Education: A Survey of the Theory and Evidence.” In *Handbook of Social Economics*, edited by Jess Benhabib, Alberto Bisin, and Matthew O. Jackson. vol. 1. 1053–1163. San Diego: North-Holland, 2011. https://doi.org/10.1016/B978-0-444-53707-2.00003-7.

Ewing, John. “Mathematical Intimidation: Driven by the Data.” *Notices of the American Mathematical Society* 58, no. 5 (May 2011): 667–73. http://www.ams.org/notices/201105/rtx110500667p.pdf.

Farson, Richard, and Ralph Keyes. “The Failure-Tolerant Leader.” *Harvard Business Review*, August 1, 2002. https://hbr.org/2002/08/the-failure-tolerant-leader.

Fischbein, Efraim, and Avikam Gazit. “Does the Teaching of Probability Improve Probabilistic Intuitions?: An Exploratory Research Study.” *Educational Studies in Mathematics* 15, no. 1 (February 1984): 1–24. https://doi.org/10.1007/BF00380436.

Fischhoff, Baruch, and Maya Bar-Hillel. “Focusing Techniques: A Shortcut to Improving Probability Judgments?.” *Organizational Behavior and Human Performance* 34, no. 2 (October 1984): 175–94. https://doi.org/10.1016/0030-5073(84)90002-3.

Forsyth, Michael, and David A. Bushey. “The Recognition-Primed Decision Model: An Alternative to the MDMP for GWOT.” *Field Artillery*, January 1, 2006.

Friedman, J., J. Baker, B. Mellers, P. Tetlock, and Richard Zeckhauser. "The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament." *International Studies Quarterly* (forthcoming) (2017): abstract. http://sites.dartmouth.edu/friedman/files/2017/06/Value-of-Precision-June-17.pdf.

Friedman, Thomas L. *Thank You for Being Late: An Optimist's Guide to Thriving in the Age of Accelerations*. New York: Farrar, Straus and Giroux, 2016. Kindle.

Friz Freleng, Warner Brothers. "Tree for Two." YouTube video, 0:44. 1952. https://www.youtube.com/watch?v=UVNHcob3oJg.

Galef, Julia. "A Visual Guide to Bayesian Thinking." YouTube video, 11:24. July 16, 2015. https://www.youtube.com/watch?v=BrK7X_XlGB8.

Gallup. "U.S. Employee Engagement 2011–2017." July 30, 2017. http://news.gallup.com/poll/214961/gallup-employee-engagement.aspx.

Galton, Francis. *Hereditary Genius: An Inquiry into Its Laws and Consequences*. 2nd ed. London: Macmillan, 1892. http://galton.org/books/hereditary-genius/text/pdf/galton-1869-genius-v3.pdf.

———. *Memories of My Life*. London: Methuen, 1908. http://galton.org/books/memories/galton-memories-1up-v2-300dpi.pdf.

———. "Sir Francis Galton F.R.S." Accessed June 3, 2018. http://galton.org/main.html.

———. Galton, Francis. "Vox Populi." *Nature* 75 (1907): 450–451. https://www.nature.com/nature/journal/v75/n1949/pdf/075450a0.pdf?foxtrotcallback=true.

Gershenfeld, Neil. "Truth is a Model." In *This Will Make You Smarter: New Scientific Concepts to Improve Your Thinking*, edited by John Brockman, 72–73. New York: Harper Perennial, 2012.

Gigerenzer, Gerd. "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases.'" *European Review of Social Psychology*, 2, no. 1 (1991): 83–115.

———. "On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky." *Psychological Review* 103, no. 3 (1996): 592–96.

GitHub. "GitHub Features: The Right Tools for the Job." Accessed August 30, 2018. https://github.com/features.

Goering, Sara. "Eugenics." In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta. Stanford, CA: Metaphysics Research Lab, Stanford University, 2014. https://plato.stanford.edu/archives/fall2014/entries/eugenics/.

Goldenkoff, Robert. *Federal Workforce: Sustained Attention to Human Capital Leading Practices Can Help Improve Agency Performance*. GAO-17-627T. Washington, DC: Government Accountability Office, 2017.

Goldin, Claudia, and Cecilia Rouse. "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians." *The American Economic Review* 90, no. 4 (2000): 715–41. http://www.jstor.org/stable/117305.

Good Judgment Project. "GJP Data." Harvard Dataverse, 2016. https://doi.org/10.7910/DVN/BPCDH5.

———. "GJP Data Year One." Harvard Dataverse, 2016. https://doi.org/10.7910/DVN/BPCDH5.

Good Judgment Open. "Frequently Asked Questions (FAQ)." Last updated June 16, 2017. https://www.gjopen.com/faq.

Government Accountability Office. *DHS Training: Improved Documentation, Resource Tracking, and Performance Measurement Could Strengthen Efforts*. GAO-14-688. Washington, DC: Government Accountability Office, 2014. http://www.gao.gov/products/GAO-14-688.

Grove, William M., David H. Zald, Boyd S. Lebow, Beth E. Snitz, and Chad Nelson. "Clinical Versus Mechanical Prediction: A Meta-Analysis." *Psychological Assessment* 12, no. 1 (2000): 19–30. https://doi.org/10.1037//1040-3590.12.1.19.

Hammond, Jeremy R. "The 'Forgotten' U.S. Shootdown of Iranian Airliner Flight 655." *Foreign Policy Journal*, July 3, 2017. https://www.foreignpolicyjournal.com/2017/07/03/the-forgotten-us-shootdown-of-iranian-airliner-flight-655%c2%ad/.

Harter, James K., Frank L. Schmidt, Sangeeta Agrawal, and Stephanie K. Plowman. *The Relationship between Engagement at Work and Organizational Outcomes*. Washington, DC: Gallup, 2016. http://www.workcompprofessionals.com/advisory/2016L5/august/MetaAnalysis_Q12_ResearchPaper_0416_v5_sz.pdf.

Haselton, Martie G., Daniel Nettle, and Paul W. Andrews. "The Evolution of Cognitive Bias." In *The Handbook of Evolutionary Psychology*, edited by David M. Buss, 724–46. Hoboken, NJ John Wiley & Sons, 2005.

Hattie, John, and H. W. Marsh. "The Relationship between Research and Teaching: A Meta-Analysis." *Review of Educational Research* 66, no. 4 (1996): 507–42. https://doi.org/10.2307/1170652.

Hausser, Doris. "Understanding the Federal Employee Viewpoint Survey." Working paper, National Academy of Public Administration, 2018. https://www.napawash.org/uploads/AWP_2_Understanding_the_Federal_Employee_Viewpoint_Survey.pdf.

Hertel, Guido, Norbert L. Kerr, and Lawrence A. Messé. "Motivation Gains in Performance Groups: Paradigmatic and Theoretical Developments on the Köhler Effect." *Journal of Personality and Social Psychology* 79, no. 4 (2000): 580–601. https://doi.org/10.1037//0022-3514.79.4.580.

Hoffman, Bruce. Twitter Post. September 9, 2018, 4:16 AM. https://twitter.com/hoffman_bruce/status/1038748067930009600.

Hogan, Robert, and Robert B. Kaiser. "What We Know about Leadership." *Review of General Psychology* 9, no. 2 (2005): 169–80. https://doi.org/10.1037/1089-2680.9.2.169.

Hogg, Michael A. "A Social Identity Theory of Leadership." *Personality and Social Psychology Review* 5, no. 3 (August 1, 2001): 184–200. https://doi.org/10.1207/S15327957PSPR0503_1.

Hong, Lu, and Scott E. Page. "Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers." *Proceedings of the National Academy of Sciences* 101, no. 46 (November 16, 2004): 16385–89. https://doi.org/10.1073/pnas.0403723101.

Hornstein, Henry A. "Student Evaluations of Teaching Are an Inadequate Assessment Tool for Evaluating Faculty Performance." Edited by Hau Fai Edmond Law, *Cogent Education* 4, no. 1 (March 20, 2017). https://doi.org/10.1080/2331186X.2017.1304016.

Howden, Daniel. "The Illusion of the Leadership Industry." Recruiting Resources: How to Recruit and Hire Better, March 31, 2016. https://resources.workable.com/blog/failures-of-leadership-industry.

Hubbard, Douglas W. *The Failure of Risk Management: Why It's Broken and How to Fix It*. 1st ed. Hoboken, NJ Wiley, 2009.

Huber, Mary, and Pat Hutchings. "New Teaching Positions up the Ante on Pedagogical Knowledge and Skill." *Bay View Alliance* (blog), February 26, 2018. http://bayviewalliance.org/new-teaching-positions-ante-pedagogical-knowledge-skill/.

Huizing, Ard, and Jan A. van der Wal. "Explaining the Rise and Fall of the Warez MP3 Scene: An Empirical Account from the Inside." *First Monday* 19, no. 10 (October 6, 2014): 1–5. http://journals.uic.edu/ojs/index.php/fm/article/view/5546.

Ignatius, David. "David Ignatius: More Chatter than Needed." *Washington Post*, sec. Opinions, November 1, 2013. https://www.washingtonpost.com/opinions/david-ignatius-more-chatter-than-needed/2013/11/01/1194a984-425a-11e3-a624-41d661b0bb78_story.html.

IMDb. "The Wire." Accessed October 7, 2018. http://www.imdb.com/title/tt0306414/.

Janis, Irving. “Groupthink.” *Psychology Today*, November 1971. http://agcommtheory.pbworks.com/f/GroupThink.pdf.

Janis, Irving L. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. 2nd ed. Boston: Houghton Mifflin, 1982.

Jarrett, Christian. “Estimating the Reproducibility of Psychological Science.” *The British Psychological Society Research Digest*, August 27, 2015. https://digest.bps.org.uk/2015/08/27/this-is-what-happened-when-psychologists-tried-to-replicate-100-previously-published-findings/.

John, Leslie K., George Loewenstein, and Drazen Prelec. “Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling.” *Psychological Science* 23, no. 5 (May 2012): 524–32. https://doi.org/10.1177/0956797611430953.

Joyce, Nola. “Can You Lead Me Now? Leading in the Complex World of Homeland Security.” Master’s thesis, Naval Postgraduate School, 2007. https://calhoun.nps.edu/bitstream/handle/10945/3286/07Sep_Joyce.pdf?sequence=1&isAllowed=y.

Joyner, Mark. “Pearson’s Law and How the New “Trackers” Feature Improve Things “Exponentially.” *Simpleology* (blog), November 16, 2011. http://www.simpleology.com/blog/2011/11/pearsons-law-and-how-the-new-trackers-feature-will-improve-things-exponentially.html.

Judge, Timothy A., and Ronald F. Piccolo. “Transformational and Transactional Leadership: A Meta-Analytic Test of Their Relative Validity.” *Journal of Applied Psychology* 89, no. 5 (2004): 755–68. https://doi.org/10.1037/0021-9010.89.5.755.

Judge, Timothy A., Joyce E. Bono, Remus Ilies, and Megan W. Gerhardt. “Personality and Leadership: A Qualitative and Quantitative Review.” *Journal of Applied Psychology* 87, no. 4 (2002): 765–80. https://doi.org/10.1037//0021-9010.87.4.765.

Kafka, Alexander C. “Why Does Publishing Higher-Ed Research Take So Long?.” *The Chronicle of Higher Education*, August 16, 2018. https://www.chronicle.com/article/Why-Does-Publishing-Higher-Ed/244291.

Kahneman, Daniel. “Don’t Blink! The Hazards of Confidence.” *New York Times*, sec. Magazine. October 19, 2011. https://www.nytimes.com/2011/10/23/magazine/dont-blink-the-hazards-of-confidence.html.

Kahneman, Daniel, and Amos Tversky. “Availability: A Heuristic for Judging Frequency and Probability.” *Cognitive Psychology* 5 (1973): 207–32. https://msu.edu/~ema/803/Ch11-JDM/2/TverskyKahneman73.pdf.

———. “Intuitive Prediction: Biases and Corrective Procedures.” June 1977. http://www.dtic.mil/docs/citations/ADA047747.

———. “On the Reality of Cognitive Illusions.” *Psychological Review* 103, no. 3 (1996): 582–91.

———. “Subjective Probability: A Judgment of Representativeness.” *Cognitive Psychology* 3 (1972): 430–54. http://datacolada.org/wp-content/uploads/2014/08/Kahneman-Tversky-1972.pdf.

Kahneman, Daniel, and Gary Klein. “Conditions for Intuitive Expertise: A Failure to Disagree.” *American Psychologist* 64, no. 6 (2009): 515–26. https://doi.org/10.1037/a0016755.

Kahneman, Daniel, and Shane Frederick. “Representativeness Revisited: Attribute Substitution in Intuitive Judgment.” In *Heuristics and Biases*, edited by Thomas Gilovich, Dale Griffin, and Daniel Kahneman, 49–81. Cambridge: Cambridge University Press, 2002. https://doi.org/10.1017/CBO9780511808098.004.

Kahneman, Daniel. “Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice.” In *Les Prix Nobel: The Nobel Prizes 2002*, edited by Tore Frangsmyr. 465–74. Stockholm: Nobel Foundation, 2003. https://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2002/kahnemann-lecture.pdf.

———. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2013. Kindle.

Kaminsky, Mark T. “Effective Selection: A Study of First-Line Supervisor Selection Processes in the Department of Homeland Security.” Master’s thesis, Naval Postgraduate School, 2011. http://www.dtic.mil/docs/citations/ADA543301.

Kirkpatrick, Shelly, and Edwin Locke. “Leadership: Do Traits Matter?.” *The Executive* 5, no. 2 (May 1991): 48–60. https://sites.fas.harvard.edu/~soc186/AssignedReadings/Kirkpatrick-Traits.pdf.

Klein, Aaron. “H.R. McMaster-Endorsed Book Calls Jihad Peaceful, Al-Qaida Terrorism ‘Resistance.’” Breitbart, August 18, 2017. https://www.breitbart.com/middle-east/2017/08/18/h-r-mcmaster-endorsed-book-calls-jihad-peaceful-al-qaida-terrorism-resistance/.

Klein, Gary A. *A Recognition-Primed Decision (RPD) Model of Rapid Decision Making*. New York: Ablex Publishing Corporation, 1993.

———. “Naturalistic Decision Making.” *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, no. 3 (June 2008): 456–60. https://doi.org/10.1518/001872008X288385.

———. *Sources of Power: How People Make Decisions*. 20th ed. Cambridge, MA: MIT Press, 2017.

Klein, Gary, Roberta Calderwood, and Anne Clinton-Cirocco. "Rapid Decision Making on the Fire Ground: The Original Study Plus a Postscript." *Journal of Cognitive Engineering and Decision Making* 4, no. 3 (September 1, 2010): 186–209. https://doi.org/10.1518/155534310X12844000801203.

KnowledgeAtWharton. "Superforecaster Full Video." YouTube video, 1:01:32. February 26, 2016. https://www.youtube.com/watch?v=6POQjSjIXWk.

Kuncel, Nathan R., Sarah A. Hezlett, and Deniz S. Ones. "Academic Performance, Career Potential, Creativity, and Job Performance: Can One Construct Predict Them All?." *Journal of Personality and Social Psychology* 86, no. 1 (2004): 148–61. https://doi.org/10.1037/0022-3514.86.1.148.

Lerner, Jennifer, and Philip Tetlock. "Accounting for the Effects of Accountability." *Psychological Bulletin* 125, no. 2 (1999): 255–75. http://scholar.harvard.edu/files/jenniferlerner/files/lerner_and_tetlock_1999_pb_paper.pdf.

Lewis, Michael. *Moneyball: The Art of Winning an Unfair Game*, 1st. pbk. ed. New York: Norton, 2004.

Lichtenstein, Sarah, Paul Slovic, Baruch Fischhoff, Mark Layman, and Barbara Combs. "Judged Frequency of Lethal Events." *Journal of Experimental Psychology: Human Learning and Memory* 4 (November 1, 1978): 551–78. https://doi.org/10.1037/0278-7393.4.6.551.

Luthans, Fred. "Successful vs. Effective Real Managers." *The Academy of Management Executive* 2, no. 2 (1987): 127–32. http://www.jstor.org/stable/4164814.

Markon, Jerry. "DHS Studies Its Endless Morale Problems, Then Studies Them Some More." *Washington Post*, sec. Politics, February 20, 2015. https://www.washingtonpost.com/politics/homeland-security-has-done-little-for-low-morale-but-study-it--repeatedly/2015/02/20/f626eba8-b15c-11e4-886b-c22184f27c35_story.html.

———. "Homeland Security Ranks Dead Last in Morale—Again—but Jeh Johnson's Morale Is High." *Washington Post*, September 29, 2015. https://www.washingtonpost.com/news/federal-eye/wp/2015/09/29/dhs-disappointed-by-latest-low-morale-scores-vows-to-keep-trying/?utm_term=.c81030c66e3b.

Marsh, Herbert W., and John Hattie. "The Relation between Research Productivity and Teaching Effectiveness: Complementary, Antagonistic, or Independent Constructs?." *The Journal of Higher Education* 73, no. 5 (2002): 603–41. https://doi.org/10.1353/jhe.2002.0047.

McCandless, David. “Warez Wars.” Wired. April 1, 1997. https://www.wired.com/1997/04/ff-warez/.

McChrystal, General Stanley, Tantum Collins, David Silverman, and Chris Fussell. *Team of Teams: New Rules of Engagement for a Complex World*. 1st ed. New York: Portfolio, 2015.

McCreary, John. “Professionalism in Analysis.” Lecture, Naval Postgraduate School, CA, December 23, 2009.

Mellers, Barbara, Eric Stone, Pavel Atanasov, Nick Rohrbaugh, S. Emlen Metz, Lyle Ungar, Michael M. Bishop, Michael Horowitz, Ed Merkle, and Philip Tetlock. “The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics.” *Journal of Experimental Psychology: Applied* 21, no. 1 (2015): 90–103. https://doi.org/10.1037/xap0000040.

Mellers, Barbara, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, and Michael Horowitz. “Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions.” *Perspectives on Psychological Science* 10, no. 3 (2015): 267–281. http://journals.sagepub.com/doi/abs/10.1177/1745691615577794.

Mellers, Barbara, Joshua Baker, Eva Chen, David Mandel, and Philip E. Tetlock. “How Generalizable Is Good Judgement? A Multi-Task, Multi-Benchmark Study.” *Judgement and Decision Making* 12, no. 4 (July 2017): 369–81. http://journal.sjdm.org/17/17408/jdm17408.pdf.

Mellers, Barbara, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, and Sydney E. Scott et al. “Psychological Strategies for Winning a Geopolitical Forecasting Tournament.” *Psychological Science* 25, no. 5 (May 2014): 1–10. https://doi.org/10.1177/0956797614524255.

Mellers, Barbara, Philip E. Tetlock, Joshua Baker, Jeffrey Friedman, and Richard Zeckhauser. “Improving the Accuracy of Geopolitical Risk Assessments.” In *The Future of Risk Management*, edited by Robert Meyer and Erwann Michel-Kerjan, 1–28. Philadelphia: University of Pennsylvania Press, forthcoming. https://sites.hks.harvard.edu/fs/rzeckhau/Geopolitical%20Risks.pdf.

Mensa International. “What Is Mensa?” Accessed September 18, 2018. https://www.mensa.org/.

Miller, Jeffrey M. *Rescuing Tomorrow Today: Fixing Training and Development for DHS Leaders*, Accession Number: AD1029855. Monterey, CA: Naval Postgraduate School, 2016. http://www.dtic.mil/docs/citations/AD1029855.

Mind Mastery. "The Complete List of Cognitive Biases." Accessed November 4, 2017. http://www.mind-mastery.com/article/322/The-Complete-List-of-Cognitive-Biases.

Mitchell, Deborah J., J. Edward Russo, and Nancy Pennington. "Back to the Future: Temporal Perspective in the Explanation of Events." *Journal of Behavioral Decision Making* 2, no. 1 (January 1, 1989): 25–38. https://doi.org/10.1002/bdm.3960020103.

Mlodinow, Leonard. "Mindware and Superforecasting." *New York Times*, October 15, 2015.

Moodle. "Moodle Higher Education." Accessed September 18, 2018. https://moodle.com/higher-education/.

Morris, Nicholas H. "A Review of Court Cases Involving Cognitive Ability Testing and Employment Practices: 1992–2015." Paper 1575. Master's thesis, Western Kentucky University, 2016. https://digitalcommons.wku.edu/cgi/viewcontent.cgi?article=2579&context=theses.

Morrissey, Ed. "We've Gotta Protect Our Phony Baloney Jobs!." YouTube video, 0:20. December 23, 2010. https://www.youtube.com/watch?v=uTmfwklFM-M.

Muller, Jerry Z. *The Tyranny of Metrics*. Princeton, Oxford: Princeton University Press, 2018. Epigraph quoting Aaron Haspel. https://www.amazon.com/Tyranny-Metrics-Jerry-Z-Muller/dp/0691174954.

National Academy of Public Administration. *Building a 21st Century SES: Ensuring Leadership Excellence in Our Federal Government*. Washington, DC: National Academy of Public Administration, 2017. https://www.napawash.org/uploads/Academy_Studies/Building-a-21st-Century-SES-3.17.2017.pdf.

Naval Postgraduate School. "FAO Asia In-Residence Course 21 June–2 July 2010—Intelligence Analysis and Professionalism Mr. John McCreary." Video, July 27, 2010. http://web.nps.edu/Video/Portal/Video.aspx?enc=MPCfFMbHCWexu6JsvpD%2FxQDYA8NNCnml.

Nicolson, Adam. *Seize the Fire: Heroism, Duty, and Nelson's Battle of Trafalgar*. New York: Harper Perennial, 2006.

Nietzsche, Friedrich Wilhelm. *Beyond Good and Evil: Prelude to a Philosophy of the Future* Translated by Walter Arnold Kaufmann. New York: Vintage Books, 1989.

Office of Personnel Management. "Unlocking Federal Talent." Accessed March 11, 2018. https://www.unlocktalent.gov/employee-engagement.

Office of the Chief Human Capital Officer. *2016 Accomplishments Report*. Washington, DC: Department of Homeland Security, 2016.

Parrish, Share. "Philip Tetlock on the Art and Science of Prediction." *The Knowledge Project*, Podcast audio. December 8, 2015.

Partnership for Public Service. "Agency Report: Department of Homeland Security." Accessed September 29, 2017. http://bestplacestowork.org/BPTW/rankings/detail/HS00.

———. "Best Places to Work Agency Rankings." Accessed March 11, 2018. http://bestplacestowork.org/BPTW/rankings/overall/large.

Pearson. "Miller Analogies Test (MAT)." Accessed September 18, 2018. https://www.pearsonassessments.com/postsecondaryeducation/graduate_admissions/mat.html.

Peter-Koop, Andrea. "Fermi Problems in Primary Mathematics Classrooms." *Australian Primary Mathematics Classroom Institute of Education Sciences* 10, no. 1 (2005): 1–5. https://files.eric.ed.gov/fulltext/EJ793997.pdf.

Pfeffer, Jeffrey. "Leadership BS: Fixing Workplaces and Careers One Truth at a Time." PowerPoint presentation, Stanford Social Innovation Review, Stanford, CA, October 2016. http://www.ssirinstitute.org/wp-content/uploads/2016/06/Pfeffer_Nonprofit-Management-Institute-2016.pdf.

Phillips, Katherine W. "How Diversity Makes Us Smarter." *Scientific American*. Accessed May 1, 2018. https://doi.org/10.1038/scientificamerican1014-42.

Pinsen, David. "Interview with a Superforecaster." Seeking Alpha, February 10, 2016. https://seekingalpha.com/article/3882906-interview-superforecaster.

Posner, Richard. *Public Intellectuals: A Study of Decline: A Critical Analysis*. Cambridge: Harvard University Press, 2001.

Pradeep, Durga Devi, and N. R. V. Prabhu. "The Relationship between Effective Leadership and Employee Performance." In *International Conference on Advancements in Information Technology with Workshop of ICBMG IPCSIT Vol. 20 IACSIT Press, Singapore, 198*, vol. 207, 2011.

Ramis, Harold, dir. *Caddyshack*. 1980; Los Angeles, CA: Orion Pictures, 1980. https://www.imdb.com/title/tt0080487/.

Rehn, Alf. "The Politics of Contraband." *The Journal of Socio-Economics* 33, no. 3 (July 2004): 359–74. https://doi.org/10.1016/j.socec.2003.12.027.

Resnick, Brian, and Julia Belluz. “A Top Cornell Food Researcher Has Had 13 Studies Retracted. That’s a Lot.” Vox, September 21, 2018. https://www.vox.com/science-and-health/2018/9/19/17879102/brian-wansink-cornell-food-brand-lab-retractions-jama.

Ricciuti, James E. “Groupthink: A Significant Threat to the Homeland Security of the United States.” Master’s thesis, Naval Postgraduate School, 2014. https://calhoun.nps.edu/bitstream/handle/10945/44650/14Dec_Ricciuti_James.pdf?sequence=1&isAllowed=y.

Robinson, Dilys, Sarah Perryman, and Sue Hayday. *The Drivers of Employee Engagement*, Report 408. Brighton, UK: Institute for Employment Studies, 2004. http://www.employment-studies.co.uk/system/files/resources/files/408.pdf.

Rock, David, and Heidi Grant. “Why Diverse Teams Are Smarter.” *Harvard Business Review*, November 4, 2016. https://hbr.org/2016/11/why-diverse-teams-are-smarter.

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, Douglas O. Staiger. *Can You Recognize an Effective Teacher When You Recruit One?* Cambridge, MA: National Bureau of Economic Research, 2011. http://www.nber.org/papers/w14485.

Roese, Neal J., and Kathleen D. Vohs. “Hindsight Bias.” *Perspectives on Psychological Science* 7, no. 5 (September 1, 2012): 411–26. https://doi.org/10.1177/1745691612454303.

Ross, Karol G., Gary A. Klein, Peter Thunholm, John F. Schmitt, and Holly C. Baxter. “The Recognition-Primed Decision Model.” *Military Review*, August 2004.

Schmidt, Frank L. “The Role of General Cognitive Ability and Job Performance: Why There Cannot Be a Debate.” *Human Performance* 15, no. 1 (2002): 187–210. https://www.researchgate.net/publication/240237107_The_Role_of_General_Cognitive_Ability_and_Job_Performance_Why_There_Cannot_Be_a_Debate.

Schmidt, Frank L., and John Hunter. “General Mental Ability in the World of Work: Occupational Attainment and Job Performance.” *Journal of Personality and Social Psychology* 86, no. 1 (2004): 162–73. https://doi.org/10.1037/0022-3514.86.1.162.

Schmidt, Frank L., and John E. Hunter. “The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings.” *Psychological Bulletin* 124, no. 2 (1998): 262–274. http://psycnet.apa.org/psycinfo/1998-10661-006.

Schoemaker, Paul, and Philip Tetlock. "Superforecasting: How to Upgrade Your Company's Judgment." *Harvard Business Review* 94 (May 2016): 12. http://mena-speakers.com/wp-content/uploads/2016/12/2016-hbr-final-final-version.pdf.

Schulz-Hardt, Stefan, and Felix C. Brodbeck. "Group Performance and Leadership." *An Introduction to Social Psychology*, 2012.

Segre, Gino. "Gedankenexperiment." *2011: What Scientific Concept Would Improve Everybody's Cognitive Toolkit?* (blog), 2011. https://www.edge.org/response-detail/10157.

Seidenfeld, Mark. "Cognitive Loafing, Social Conformity and Judicial Review of Agency Rulemaking." *Cornell Law Review* 87, no. 2 (January 2002): 511–12. https://doi.org/10.2139/ssrn.280251.

Serio, Tricia. "Opinion: Repairing Peer Review." *The Scientist Magazine*®, November 18, 2016. https://www.the-scientist.com/opinion/opinion-repairing-peer-review-32512.

Shermer, Michael. "Patternicity: Finding Meaningful Patterns in Meaningless Noise." *Scientific American* 299, no. 48 (2008). https://www.scientificamerican.com/article/patternicity-finding-meaningful-patterns/.

Shostak, Seth. "What Scientific Term or Concept Ought to Be More Widely Known? Fermi Problems." Edge. Accessed August 15, 2018. https://www.edge.org/response-detail/27055.

Slovic, Paul, and Baruch Fischhoff. "On the Psychology of Experimental Surprises." *Journal of Experimental Psychology: Human Perception and Performance* 3, no. 4 (1977): 544–51. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.365.7695&rep=rep1&type=pdf.

Society for Industrial and Organizational Psychology, Inc. "Effective Interviews." Accessed September 18, 2018. http://www.siop.org/workplace/employment%20testing/interviews.aspx.

———. "Welcome to SIOP." Accessed September 18, 2018. https://www.siop.org.

Somesh, K. S. "The Greatest Enemy of Knowledge Is Not Ignorance, It Is the Illusion of Knowledge." *Medium* (blog), March 8, 2018. https://medium.com/@ks.somesh 2016/the-greatest-enemy-of-knowledge-is-not-ignorance-it-is-the-illusion-of-knowledge-5c0dd1dcca7e quoting Neil Boorstin.

Sontag, Sherry, Christopher Drew, and Annette Lawrence Drew. *Blind Man's Bluff: The Untold Story of American Submarine Espionage*. New York: PublicAffairs, 2016.

Stewart, Jon, *Daily Show*. Comedy Central, March 9, 2009.

Stewart, Thomas R. "Improving Reliability of Judgmental Forecasts." In *Principles of Forecasting*, edited by J. Scott Armstrong, 81–106. vol. 30. Boston: Springer U.S., 2001. https://doi.org/10.1007/978-0-306-47630-3_5.

Stutzman, Mike, and Tracy K. Tunwall. "Leadership Success or Failure: Understanding the Link between Promotion Criteria and Leader Effectiveness." *Journal of Business and Economics* 4, no. 8 (August 2003): 690–94. http://173.83.167.93/UploadFile/Picture/2014-6/201461493432288.pdf.

Surowiecki, James. *The Wisdom of Crowds*. 1st ed. New York: Anchor Books, 2005.

Taggar, Simon, Rick Hackew, and Sudhir Saha. "Leadership Emergence in Autonomous Work Teams: Antecedents and Outcomes." *Personnel Psychology* 52, no. 4 (December 1999): 899–926. https://doi.org/10.1111/j.1744-6570.1999.tb00184.x.

Tajfel, Henry, and John Turner. "An Integrative Theory of Intergroup Conflict." In *The Social Psychology of Intergroup Relations*, edited by William Austin and Stephen Worchel, ch. 3, 33–47. Monterey, CA: Brooks/Cole, 1979.

Taleb, Nassim Nicholas. *The Black Swan: The Impact of the Highly Improbable*. 2nd ed. Random House Trade pbk. ed. New York: Random House Trade Paperbacks, 2010.

Tetlock, Philip. "Edge Master Class 2015: A Short Course in Superforecasting, Class II." Edge, August 24, 2015. https://www.edge.org/conversation/philip_tetlock-edge-master-class-2015-a-short-course-in-superforecasting-class-ii.

———. *Expert Political Judgement: How Good is It? How Can We Know?* Princeton, NJ: Princeton University Press, 2005.

———. *Expert Political Judgment: How Good Is It? How Can We Know?* 1st. pbk. ed. Princeton, NJ: Princeton University Press, 2006.

———. "How to Win at Forecasting: A Conversation with Philip Tetlock." Edge. Accessed August 3, 2018. https://www.edge.org/conversation/philip_tetlock-how-to-win-at-forecasting.

Tetlock, Philip, and Dan Gardner. "We Can Learn to Predict Future Events." *The Telegraph*, October 30, 2015, 1. http://www.telegraph.co.uk/news/uknews/defence/11965831/We-can-learn-to-predict-future-events.html.

Tetlock. Philip E. "Accountability: A Social Check on the Fundamental Attribution Error." *Social Psychology Quarterly* 48, no. 3 (1985): 227–36. https://doi.org/10.2307/3033683.

Tetlock, Philip E., and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. New York: Random House, 2015.

Tetlock, Philip E., Barbara A. Mellers, Nick Rohrbaugh, and Eva Chen. "Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate." *Current Directions in Psychological Science* 23, no. 4 (August 2014): 290–95. https://doi.org/10.1177/0963721414534257.

Tolkien, J.R.R. *The Fellowship of the Ring*. New York: Ballantine Books, 1977.

Tversky, Amos, and Daniel Kahneman. "Belief in the Law of Small Numbers." *Psychological Bulletin* 76, no. 2 (1971): 105–10. http://stats.org.uk/statistical-inference/TverskyKahneman1971.pdf.

———. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185, no. 4157 (1974): 1124–31. http://www.jstor.org/stable/1738360.

U.S. Coast Guard. "DHS Leader Development Program | Office of Leadership (CG-12C)." Accessed March 11, 2018. http://www.dcms.uscg.mil/Our-Organization/Assistant-Commandant-for-Human-Resources-CG-1/Civilian-Human-Resources-Diversity-and-Leadership-Directorate-CG-12/Office-of-Leadership-CG-12C/DHS-Leader-Development/.

U.S. Merit Systems Protection Board. *Call to Action: Improving First-Level Supervision of Federal Employees*. Washington, DC: U.S. Merit Systems Protection Board, 2010. https://www.mspb.gov/MSPBSEARCH/viewdocs.aspx?docnumber=516534&version=517986&application=ACROBAT.

U.S. Office of Personnel Management. "About the Federal Employee Viewpoint Survey." Accessed March 9, 2018. https://www.opm.gov/fevs/about.

U.S. Office of the Inspector General. *Major Management and Performance Challenges Facing the Department of Homeland Security*. Washington, DC: U.S. Office of the Inspector General, 2016. https://www.oig.dhs.gov/sites/default/files/assets/2017/OIG-17-08-Nov16.pdf.

University of British Columbia. "Bayes' Rule." Accessed August 13, 2018. https://www.cs.ubc.ca/~murphyk/Bayes/bayesrule.html.

Uttl, Bob, Carmela A. White, and Daniela Wong Gonzalez. "Meta-Analysis of Faculty's Teaching Effectiveness: Student Evaluation of Teaching Ratings and Student Learning Are Not Related." *Studies in Educational Evaluation* 54 (September 2017): 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007.

Vroom, Victor H., and Arthur G. Jago. "The Role of the Situation in Leadership." *American Psychologist* 62, no. 1 (2007): 17–24. https://doi.org/10.1037/0003-066X.62.1.17.

Weldon, Elizabeth. “Cognitive Loafing: The Effects of Accountability and Shared Responsibility on Cognitive Effort.” *Personality and Social Psychology Bulletin* 14, no. 1 (March 1988): 159–71. http://journals.sagepub.com/doi/pdf/10.1177/0146167288141016.

Welton, Chang, Pavel Atanasov, Shefali Patil, Barbara A. Mellers, and Philip E. Tetlock. “Accountability and Adaptive Performance under Uncertainty: A Long-Term View.” *Judgment and Decision Making* 12, no. 6 (2017): 610–626.

Wheelan, Charles J. *Naked Statistics: Stripping the Dread from the Data*. 1st. publ. as a Norton pbk. ed. New York: Norton, 2014.

White, Ralph, and Ronald Lippitt. “Leader Behavior and Member Reaction in Three ‘Social Climates.’” In *Group Dynamics: Research and Theory*, edited by Dorwin Philip Cartwright and Alvin Frederick Zander, 318–35. 3rd ed. New York: Harper & Row. 1976.https://is.muni.cz/el/1451/podzim2013/np2270/um/cartwright_leader0001.pdf.

Whitson, Jennifer A., and Adam D. Galinsky. “Lacking Control Increases Illusory Pattern Perception.” *Science* 322, no. 5898 (October 3, 2008): 115–17. https://doi.org/10.1126/science.1159845.

Worline, Monica, and Dennis Matthies. *Stanford Learning Lab Learning Careers Project: A Self-Coaching Focus*. Stanford, CA: Stanford Center for Innovative Learning, n.d. http://scil.stanford.edu/research/learningcareers/documents/selfcoach1.pdf.

Yagoda, Ben. “Your Lying Mind: The Cognitive Biases Tricking Your Brain.” *The Atlantic*, September 2018. https://www.theatlantic.com/magazine/archive/2018/09/cognitive-bias/565775/.

Yale. “Center for Teaching and Learning.” Accessed September 15, 2018. https://ctl.yale.edu/.

Yong, Ed. “A Failed Replication Draws a Scathing Personal Attack from a Psychology Professor.” Discover Magazine, *Not Exactly Rocket Science* (blog). March 10, 2012. http://blogs.discovermagazine.com/notrocketscience/2012/03/10/failed-replication-bargh-psychology-study-doyen/.

Yoon, Yeosun, Gülen Sarial-Abi, and Zeynep Gürhan-Canli. “Effect of Regulatory Focus on Selective Information Processing.” *Journal of Consumer Research* 39, no. 1 (June 1, 2012): 93–110. https://doi.org/10.1086/661935.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California