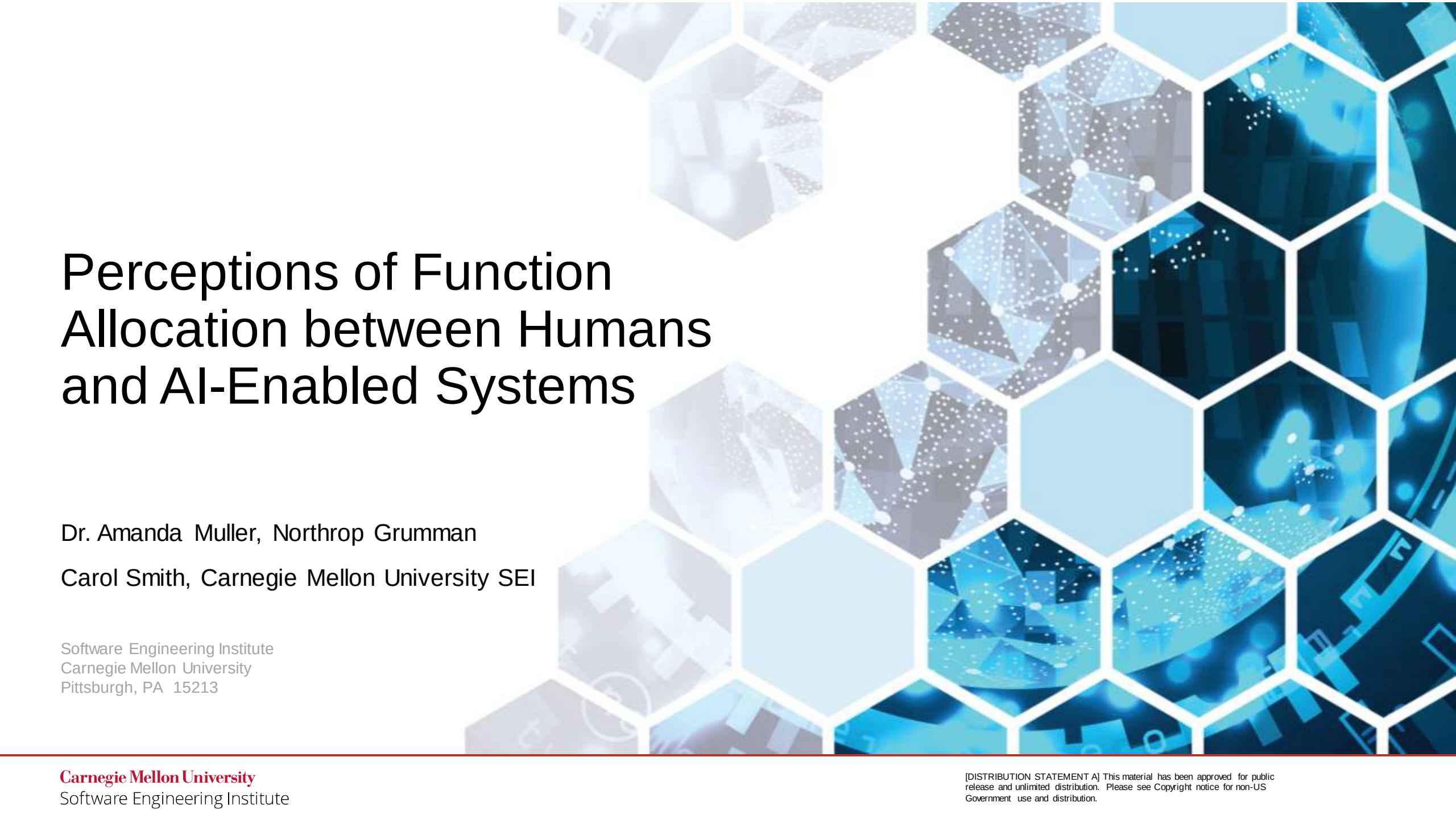




Perceptions of Function Allocation between Humans and AI-Enabled Systems

Dr. Amanda Muller, Northrop Grumman

Carol Smith, Carnegie Mellon University SEI

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Copyright Statement

Copyright 2022 Carnegie Mellon University and Northrop Grumman.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon® is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM22-0389

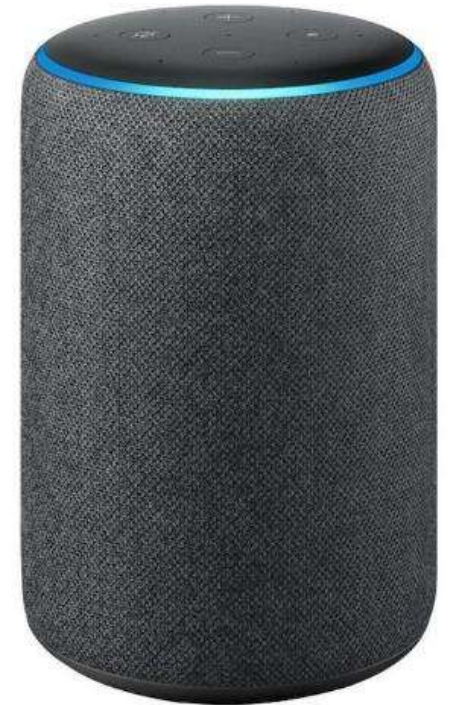
AI and UX



AI is technology that can:

- recognize patterns,
 - create predictions,
 - make decisions, and/or
 - generate new content,
- without being explicitly programmed to do so.

Examples of AI enabled systems



Machine Learning

- Typical method to create artificial intelligence (AI)
 - Pre-existing data
 - Math (statistics, probability, calculus)
 - Programming
- **Math + Programming = Algorithm (or a model)**
- **Algorithm + Data = Model**

We need to make AI human-centered



Human-Centered AI

User Experience Honeycomb

Peter Morville, et al.

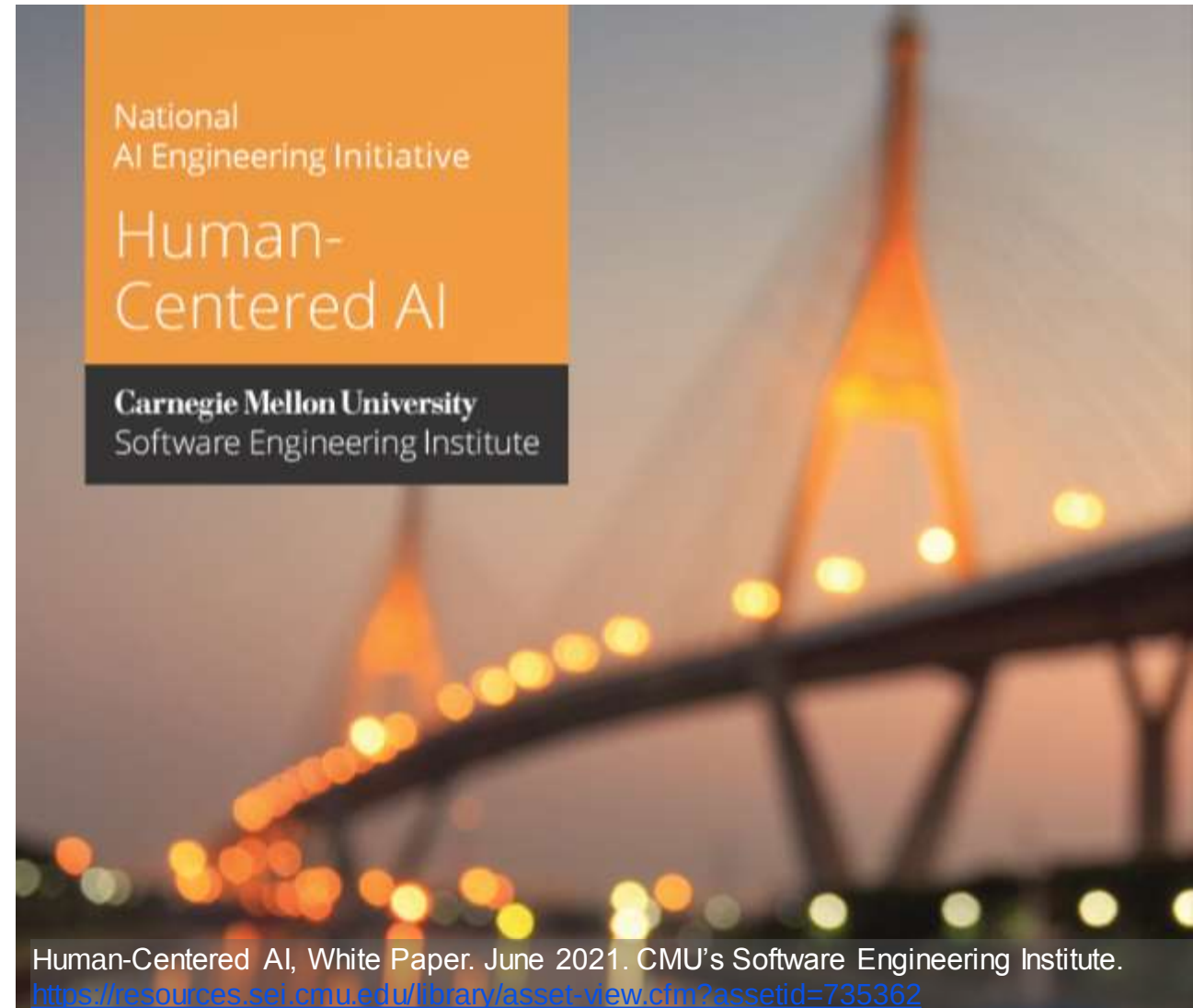
Used with permission.

Broaden our Work

- Is this an AI-friendly challenge?
- What kind of improvements are expected?
- What are the benefits and risks?
- How will we know we've made improvements?

Design to work with, and for, people

- Demystify AI
 - Minimize unintended consequences
1. Understand complexity of context
 2. Design for human-machine teaming
 3. Engage in critical oversight



The Fitts List

70 years ago

1951 Paul Fitts and team:

“surveyed the kinds of things men can do better than present-day machines, and vice versa”

Although considerable talent and skill have gone into the preparation of the report, it is nevertheless not anticipated that there will be complete agreement with the conclusions reached from the survey of past research and experience. So, for example, questions may be raised concerning the viewpoint that men are extremely poor monitors of machines, and that emphasis might well be placed on the development of systems in which the essential responsibility for monitoring would be placed upon machines. However, the very fact that the issue is raised is, in itself, of importance, particularly since research procedures are available by means of which this basic question can be settled, leading to the establishment of fundamental principles concerning the respective roles of machines and men as monitors on factual grounds.

Fitts, P. M. (Ed.) (1951). *Human engineering for an effective air-navigation and traffic-control system*. Washington, DC: National Research Council.

Fitts Original List – 11 Statements

Humans appear to surpass present-day machines in respect to the following:

1. Ability to detect a small amount of visual or acoustic energy
2. Ability to perceive patterns of light or sound
3. Ability to improvise and use flexible procedures
4. Ability to store very large amounts of information for long periods and to recall relevant facts at the appropriate time
5. Ability to reason inductively
6. Ability to exercise judgment

Present-day **machines** appear to surpass humans in respect to the following:

7. Ability to respond quickly to control signals and to apply great force smoothly and precisely
8. Ability to perform repetitive, routine tasks
9. Ability to store information briefly and then to erase it completely
10. Ability to reason deductively, including computational ability
11. Ability to handle highly complex operations, i.e. to do many different things at once

Fitts, P. M. (Ed.) (1951). *Human engineering for an effective air-navigation and traffic-control system*. Washington, DC: National Research Council.

Our Study

Overview

Due to significant advances in the field of AI we suspect that there has been a *shift in perception* of the relative proficiency of humans and AI-enabled systems abilities.

- MABA-MABA is based on old technology and outdated assumptions
- New guidance needed in the context of AI
- Conducted a literature review
- Created our list

Hypothesis: Humans are Better At

- Identifying downstream impacts
- Induction
- Judgment
- Recognizing Bias
- Responding to change
- Socio-political nuance
- Taking context into consideration

Hypothesis: AI is Better At

- Computation
- Computational Complexity
- Deduction
- Exposing Bias
- Repetition
- Replication
- Scale
- Short and long-term memory
- Simultaneous operation
- Velocity
- With context:
 - Induction
 - Judgment

Methodology

- Survey via Qualtrics
 - CMU Institutional Review Board approved
 - Validated initial statement wording with colleagues and conducted pilot study.
 - Statements presented alphabetically to reduce bias in responses.
- Recruited via email invitations
 - Self-identified as working with complex AI-enabled systems.
 - 46 respondents
 - Anonymous and no personally identifiable information was collected

Instructions – Personal assessment

Select the option corresponding to how much you believe AI surpasses humans for each function, and vice-versa.

Example Statements:

- Computational Complexity: Ability to convert energy to perform complex mathematical calculations with accuracy and precision
- Exposing Bias: Ability to identify when data or algorithms demonstrate partiality toward or against a particular conclusion
- Induction: Looking across specific observations and making broad generalizations from them. Inductive reasoning is a powerful method to analyze situations and determine what patterns are there.

All 17 statements are in the Appendix

Response choices

- Humans far surpass AI
- Humans somewhat surpass AI
- Humans and AI are equal
- AI somewhat surpasses humans
- AI far surpasses humans
- Don't know/
prefer not to answer

1. Computation: Performing complex mathematical calculations with accuracy and precision.

Humans far surpass AI ☐	Humans somewhat surpass AI ☐	Humans and AI are equal ☐	AI somewhat surpasses humans ☐	AI far surpasses humans ☐	Don't know/prefer not to answer ☐
---	--	---	--	--	---

2. Computational Complexity: Ability to convert energy to perform complex mathematical calculations with accuracy and precision.

Humans far surpass AI ☐	Humans somewhat surpass AI ☐	Humans and AI are equal ☐	AI somewhat surpasses humans ☐	AI far surpasses humans ☐	Don't know/prefer not to answer ☐
---	--	---	--	--	---

Findings and Conversation

Overall Survey Responses

Humans surpass AI*

- **Exposing Bias**
- Identifying downstream impacts
- **Judgment**
- Recognizing Bias
- Responding to change
- Socio-political nuance
- Taking context into consideration

AI surpass humans*

- Computation
- **Computational complexity**
- Repetition
- Replication
- Scale
- Short and long-term memory
- Simultaneous operation
- Velocity

*Columns show responses of “far surpass” and “somewhat surpass” combined and are significant via t-test.
Bolded items were counter to our hypothesis.

Why?

- Alphabetized list order:
 - Computation
 - Computational complexity
- Statement wording
- What do you think?

Exposing Bias

Statement: Ability to identify when data or algorithms demonstrate partiality toward or against a particular conclusion.

Survey response: Humans surpass AI

Discussion

- Wording: “Data or algorithms” - indicates computers?
- Respondents: Primarily Engineers?
- Were we wrong?

Judgement

Statement: Determining the best option - particularly when none of the options are ideal. Comparing, evaluating potential outcomes and looking at all mitigating issues. Using broad information sources when considering alternative courses of action.

Survey response: Humans surpass AI

Discussion

- Assumption: With context AI could be better (not specified in statement)
- Future work: Reliance - AI judgement vs. human judgement.
- Example use cases – considerations of AI vs. Human judgements?

Computational complexity

Statement: Ability to convert energy to perform complex mathematical calculations with accuracy and precision.

Survey response: AI far surpasses humans

- 8 survey responses: Humans far surpass AI

Discussion

- Wording: “Convert energy”
- Another issue?

Deduction and Induction

Statements

- Deduction: Examining broad data sets and making specific conclusions from them.
- Induction: Looking across specific observations and making broad generalizations from them. Inductive reasoning is a powerful method to analyze situations and determine what patterns are there.

Survey response: No statistically significant finding

Discussion

- Are humans and machines equal now?
- Wording improvements?

Appendix

Fitts Original List – 11 Statements

Humans appear to surpass present-day machines in respect to the following:

1. Ability to detect a small amount of visual or acoustic energy
2. Ability to perceive patterns of light or sound
3. Ability to improvise and use flexible procedures
4. Ability to store very large amounts of information for long periods and to recall relevant facts at the appropriate time
5. Ability to reason inductively
6. Ability to exercise judgment

Present-day **machines** appear to surpass humans in respect to the following:

7. Ability to respond quickly to control signals and to apply great force smoothly and precisely
8. Ability to perform repetitive, routine tasks
9. Ability to store information briefly and then to erase it completely
10. Ability to reason deductively, including computational ability
11. Ability to handle highly complex operations, i.e. to do many different things at once

Fitts, P. M. (Ed.) (1951). *Human engineering for an effective air-navigation and traffic-control system*. Washington, DC: National Research Council.

Fitts Original List – Comparison with Study

Humans appear to surpass present-day machines in respect to the following:

1. Ability to detect a small amount of visual or acoustic energy
2. Ability to perceive patterns of light or sound
3. Ability to improvise and use flexible procedures
4. Ability to store very large amounts of information for long periods and to recall relevant facts at the appropriate time
5. Ability to reason inductively
6. Ability to exercise judgment

Present-day **machines** appear to surpass humans in respect to the following:

7. Ability to respond quickly to control signals and to apply great force smoothly and precisely
8. Ability to perform repetitive, routine tasks
9. Ability to store information briefly and then to erase it completely
10. Ability to reason deductively, including computational ability
11. Ability to handle highly complex operations, i.e. to do many different things at once

Fitts, P. M. (Ed.) (1951). *Human engineering for an effective air-navigation and traffic-control system*. Washington, DC: National Research Council.

Computational complexity

Ability to convert energy to perform complex mathematical calculations with accuracy and precision.

Survey Response	%	Count
Humans far surpass AI	16.67%	8
Humans somewhat surpass AI	12.50%	6
Humans and AI are equal	0.00%	0
AI somewhat surpasses humans	12.50%	6
AI far surpasses humans	50.00%	24
Don't know/prefer not to answer	8.33%	4
One-sample t-test 0.000335014 Significant at p = 0.05 Significant at p=0.1	Total 100%	48

Deduction

Examining broad data sets and making specific conclusions from them.

Survey Response	%	Count
Humans far surpass AI	8.33%	4
Humans somewhat surpass AI	22.92%	11
Humans and AI are equal	12.50%	6
AI somewhat surpasses humans	33.33%	16
AI far surpasses humans	22.92%	11
Don't know/prefer not to answer	0.00%	0
Total	100%	48

One-sample t-test 0.051184831

Not significant at $p = 0.05$

Significant at $p=0.1$

Induction

Looking across specific observations and making broad generalizations from them. Inductive reasoning is a powerful method to analyze situations and determine what patterns are there.

Survey Response	%	Count
Humans far surpass AI	31.25%	15
Humans somewhat surpass AI	29.17%	14
Humans and AI are equal	8.33%	4
AI somewhat surpasses humans	12.50%	6
AI far surpasses humans	16.67%	8
Don't know/prefer not to answer	2.08%	1
	Total 100%	48
One-sample t-test 0.0511273		
Not significant at p = 0.05		
Significant at p=0.1		

Exposing Bias

Ability to identify when data or algorithms demonstrate partiality toward or against a particular conclusion.

Survey Response	%	Count
Humans far surpass AI	33.33%	16
Humans somewhat surpass AI	33.33%	16
Humans and AI are equal	8.33%	4
AI somewhat surpasses humans	10.42%	5
AI far surpasses humans	14.58%	7
Don't know/prefer not to answer	0.00%	0
	Total 100%	48
One-sample t-test 0.00674659		
Significant at p = 0.05		
Significant at p=0.1		

Judgement

Determining the best option - particularly when none of the options are ideal. Comparing, evaluating potential outcomes and looking at all mitigating issues. Using broad information sources when considering alternative courses of action.

Survey Response	%	Count
Humans far surpass AI	39.58%	19
Humans somewhat surpass AI	25.00%	12
Humans and AI are equal	14.58%	7
AI somewhat surpasses humans	10.42%	5
AI far surpasses humans	10.42%	5
Don't know/prefer not to answer	0.00%	0
One-sample t-test 0.000781279 Significant at $p = 0.05$ Significant at $p=0.1$	Total 100%	48

Definitions (1-6)

1. Computation: Performing complex mathematical calculations with accuracy and precision
2. Computational Complexity: Ability to convert energy to perform complex mathematical calculations with accuracy and precision
3. Deduction: Examining broad data sets and making specific conclusions from them.
4. Exposing Bias: Ability to identify when data or algorithms demonstrate partiality toward or against a particular conclusion
5. Identifying downstream impacts: Evaluating potential unexpected/undesirable outcomes and looking at adjacent or completely different examples and bringing those examples into the planning effort.
6. Induction: Looking across specific observations and making broad generalizations from them. Inductive reasoning is a powerful method to analyze situations and determine what patterns are there.

Definitions (7-12)

- 7. Judgment: Determining the best option – particularly when none of the options are ideal. Comparing, evaluating potential outcomes and looking at all mitigating issues. Using broad information sources when considering alternative courses of action.
- 8. Recognizing Bias: Ability to identify when a particular conclusion is partial to or against a particular type of people, data, situations, etc. group, data type, situation, etc.
- 9. Repetition: Doing the same repeatedly ad infinitum in the same way
- 10. Replication: Doing the same thing multiple times and getting the same result
- 11. Responding to change: Detecting change, integrating that information, determining a response, and responding. Improvising. Adapting and adjusting quickly to new circumstances. Ability to use existing tools and materials in new and creative ways as needed.
- 12. Scale: Applying operations on increasingly large data sets.

Definitions (13-17)

- 13. Short and long-term memory: Holding information in storage for later retrieval with accuracy and precision
- 14. Simultaneous operation: Doing multiple things at the same time with the same level of effectiveness as doing them one at a time
- 15. Socio-political nuance: Applying knowledge of historical social and political events and the influence they have on potentially unrelated topics and industries.
- 16. Taking context into consideration: Using all available signals in the environment to explore and understand a situation. The information available regarding the physical environment (other people present people, objects, etc.), the condition of the environment (safety, visibility, etc.), the perception/interpretation of that environment all contribute to context.
- 17. Velocity: Ability to perform large magnitudes of operations per second

Consent

- Purpose: The purpose of the research is to explore and compare the current perceptions of the abilities of humans and AI-enabled systems.
- Procedures: The survey is being administered using Qualtrics (this application) within your browser. In the following pages, if you choose to participate you will be asked a series of questions. We ask that you answer as best you can. The survey will take approximately 5-10 minutes to complete.
- Participant Requirements: Participation in this study is limited to individuals age 18 and older who are fluent in English. You must also self-identify as a working professional who works on complex AI-enabled systems (examples include, but are not limited to data science, machine learning, robotics, human-computer interaction, and AI software systems). You cannot be a U.S. government employee nor an active member of the US military.
- Risks: The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life or during other online activities.
- Benefits There may be no personal benefit from your participation in the study but the knowledge received may be of value to humanity.
- Compensation & Costs: There is no compensation for participation in this study. There will be no cost to you if you participate in this study.
- More details in CMU Institutional Review Board approved IRB00000603

Annotated Bibliography

Viewpoint

Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians

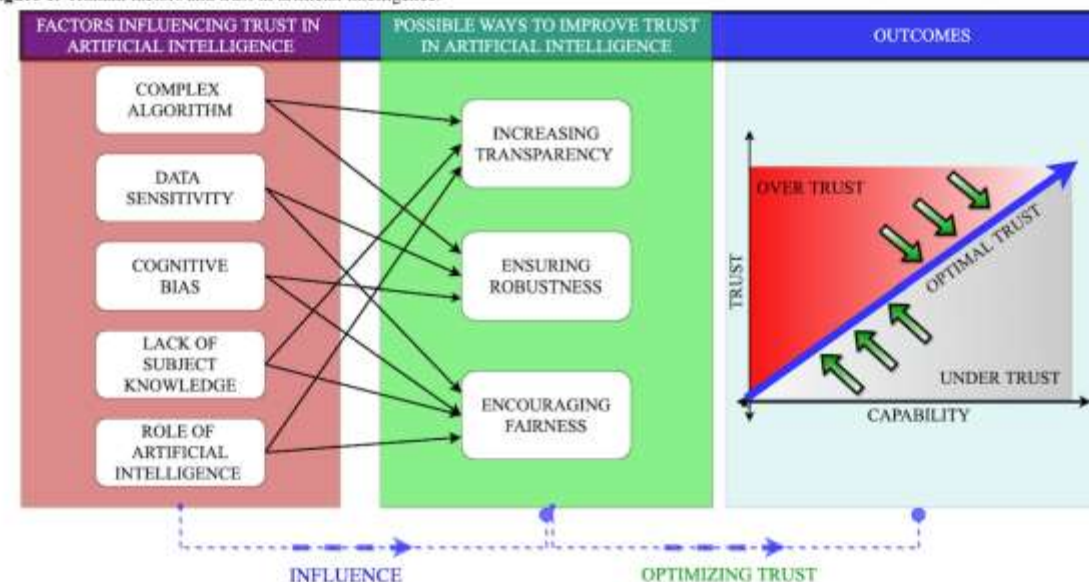
Omur Asan, PhD; Alparslan Emrah Bayrak, PhD; Avishek Choudhury, MSc
School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ, United States

Corresponding Author:

Omur Asan, PhD
School of Systems and Enterprises
Stevens Institute of Technology
1 Castle Terrace
Hoboken, NJ, 07030
United States
Phone: 1 2012163666
Email: osan@stevens.edu

- The role of humans in the practical applications of AI is often overlooked.
- Several studies have shown that automated systems may or may not improve human decision-making, depending on whether or how human factors are accounted for in their design
- AI in health care has two potential advantages to human performance.
 - First, AI can learn from big data (such as incommunicable silos of unstructured information stored in an electronic health record) more efficiently than clinicians...
 - Second, AI systems can perform predefined tasks with higher precision. AI can be in a continuous active state without compromising its performance—it does not suffer from burnout as humans do.
- There are limitations of AI that might restrict its application in life-critical areas such as healthcare.
 - AI methods require data inputs to be in a structured form, which limits the type of information that can be provided for medical decisions...
 - AI methods generally lack “common sense,” making them unable to identify simple mistakes in data or decisions that would otherwise be obvious to a human being.
- Therefore, there is significant potential and need for improvement by combining the intuitive and analytical thinking of medical experts with the computational power of AI in a proper human-AI collaboration architecture.
- [U]nnecessarily high trust in AI may have catastrophic consequences, especially in life-critical applications. Therefore, our study supports the concept of *optimal trust* in which both humans and AI each have some level of skepticism regarding the other’s decisions since both are capable of making mistakes.
- Bias, a mismatch between the distribution of training data and a preferred fair distribution, can yield unfair outcomes (prediction/classification).

Figure 1. Human factors and trust in artificial intelligence.



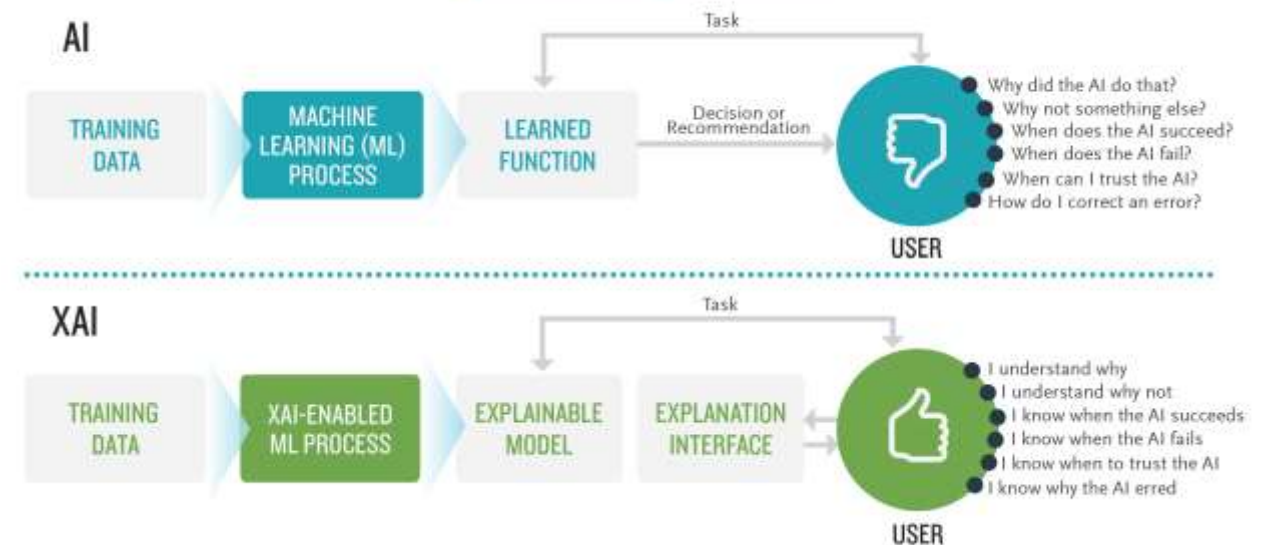
complementary capabilities of humans and automated systems [3]. The Fitts list includes 11 statements asserting that humans are better at detection, perception, judgment, induction, improvisation, and long-term memory, while automated systems are better at power/speed, computation, replication, simultaneous operation, and short-term memory [4]. Several studies have shown that automated systems may or may not improve human decision-making, depending on whether or how human factors are accounted for in their design [5].

and future research directions in the health care domain. While the users of AI systems can be diverse, including patients and insurance providers, the focus of this paper is limited to the domain experts in healthcare, ie. clinicians. We acknowledge that trust relationships could significantly differ for patients and insurance providers.

- XAI aims to create a suite of ML techniques that produce explainable model results while maintaining a high level of learning performance (or prediction accuracy).
- XAI explains decisions for five main purposes:
 - 1. Analyst Insight...
 - 2. Bias Detection & Bias Mitigation...
 - 3. Continuous Improvement...
 - 4. Quality Assurance...
 - 5. Regulatory Transparency...
- Measuring explanation effectiveness depends on the following approaches toward generating explanations:
 - Correctability...
 - Mental Model...
 - Task Performance...
 - Trust Assessment...
 - User Satisfaction...
- XAI Models
 - Visual Explanations
 - Interpretable models
 - Model induction
- Typical XAI Explainability Offerings
 - Explainability Models
 - Human in the Loop
 - Detection



Figure 1: AI and XAI Explanation Framework



- We believe thinking in this domain needs to be realigned, moving away from the traditional microergonomics conceptualisation, concerned predominantly with task-based decisions, and towards a macroergonomics approach, viewing function allocation choices as central to effective systems design.
- We believe that allocating functions in the traditional ways outlined above, so as to capitalise on the strengths of technology whilst compensating for (or eliminating) the weaknesses of humans, merely serves to create unanticipated problems.
- Thus, through automating processes, a system's capacity for adaptation may actually be reduced, increasing the inherent risks which automation was intended to remove.
- Furthermore, a key disadvantage of automating tasks previously allocated to humans is the inability of technology to take into consideration the multiple, and often conflicting, goals which frequently arise in real work settings
- We view function allocation as the process of allocating functions or tasks between and amongst humans and technologies in a system, which forms an explicit stage early in systems design and development, involves a multi-disciplinary team (including end-users), is evidence-based, framed within the language of risk, and utilises iterative methods (e.g. scenarios planning techniques).

Ergonomics, 2013

Vol. 56, No. 7, 1051–1069, <http://dx.doi.org/10.1080/00140139.2013.790482>



Function allocation in complex systems: reframing an old problem

Rose Challenger^{a*}, Chris W. Clegg^b and Craig Shepherd^c

^a*Institute of Work Psychology, Management School, The University of Sheffield, Sheffield S10 2TN, UK;* ^b*Socio-Technical Centre, Leeds University Business School, University of Leeds, Leeds LS2 9JT, UK;* ^c*Nottingham University Business School, Jubilee Campus, Nottingham NG8 1BB, UK*

(Received 6 February 2012; final version received 24 March 2013)

In this article, we offer a new, macroergonomics perspective on the long-debated issue of function allocation. We believe thinking in this domain needs to be realigned, moving away from the traditional microergonomics conceptualisation, concerned predominantly with task-based decisions, and towards a macroergonomics approach, viewing function allocation choices as central to effective systems design. We frame our arguments within a systems perspective, advocating that function allocation issues need to be on the agenda of all individuals with a wider interest in the human and organisational aspects of complex work systems, including people who commission, sponsor, design, implement and use such systems. We also argue that allocation decisions should form a transparent, explicit stage early in the systems design and development process, involve multiple stakeholders (including end-users), be evidence-based, framed within the language of risk and utilise iterative methods (e.g. scenarios planning techniques).

Practitioner Summary: This article presents a macroergonomics approach to function allocation, advocating its importance in effective systems design. Adopting a systems mindset, we argue function allocation should form an explicit stage early in the design process, involve multiple stakeholders, be evidence-based, framed within the language of risk and utilise iterative methods.

Keywords: allocation of function; complex systems; socio-technical systems; human-machine systems

- Capitalizing on some strength of automation does not replace a human weakness. It creates new human strengths and weaknesses—often in unanticipated ways
- Where automation is introduced, new human roles emerge.
- The key for designers is to accept that automation *will* transform people's practice and to be prepared to learn from these transformations as they happen.
- To be observable, automation activities should be presented in ways that capitalize on well-documented human strengths (our perceptual system's acuity to contrast, change and events; our ability to recognize patterns and know how to act on the basis of this recognition. For example:
 - Event-based...;
 - Future-oriented: ...;
 - Pattern-based:....
- Human operators must be allowed to preserve their strategic role in managing system resources as they see fit given the circumstances.



Original Article | Published: November 2002

MABA-MABA or Abracadabra? Progress on Human–Automation Co-ordination

[S. W. A. Dekker](#) & [D. D. Woods](#)

[Cognition, Technology & Work](#) 4, 240–244 (2002) | [Cite this article](#)

1234 Accesses | 160 Citations | 3 Altmetric | [Metrics](#)

Abstract:

In this paper we argue that substitution-based function allocation methods (such as MABA-MABA, or Men-Are-Better-At/Machines-Are-Better-At lists) cannot provide progress on human–automation co-ordination. Quantitative 'who does what' allocation does not work because the real effects of automation are qualitative: it transforms human practice and forces people to adapt their skills and routines. Rather than re-inventing or refining substitution-based methods, we propose that the more pressing question on human–automation co-ordination is 'How do we make them get along together?'

- Survey of changes in belief of Fitts List (as of 2015)



Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 3 (2015) 5334 – 5341

Procedia
MANUFACTURING

6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015

Reflections on the 1951 Fitts list: Do humans believe now that machines surpass them?

J.C.F. de Winter^{a,*}, P.A. Hancock^b

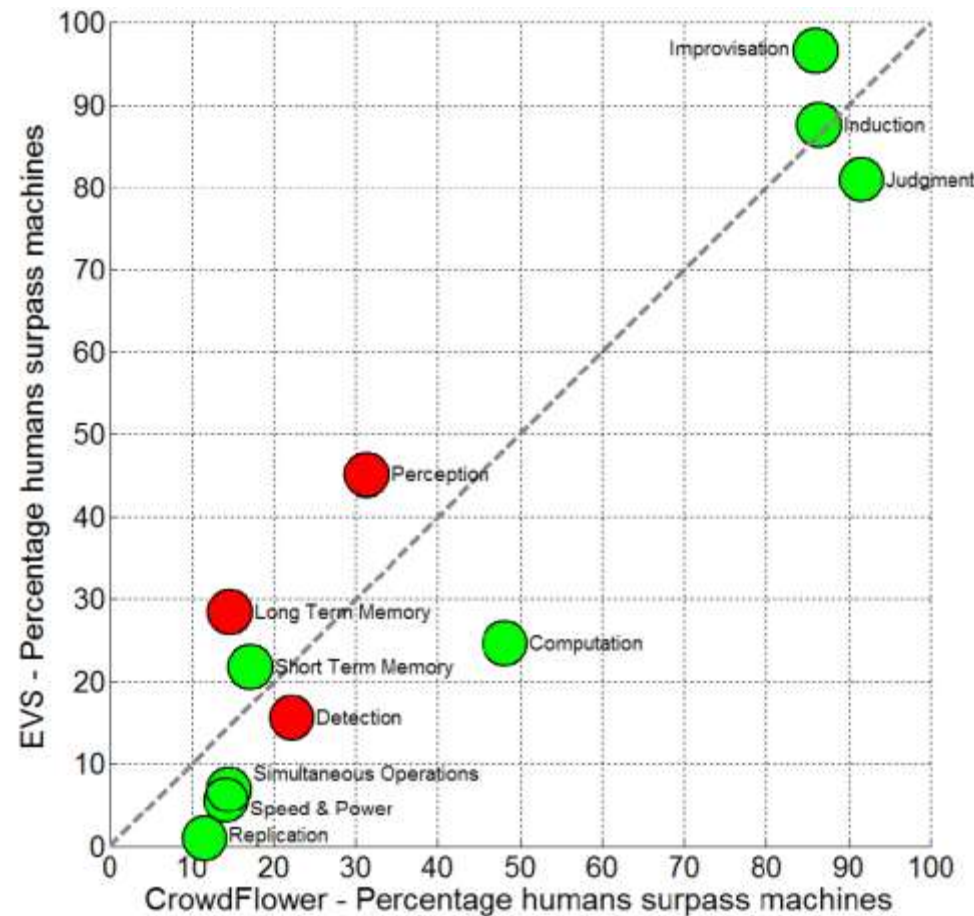


Fig. 2. Percentage stating that humans surpass machines. A red colored circle designates that the result is not in agreement with the Fitts list.

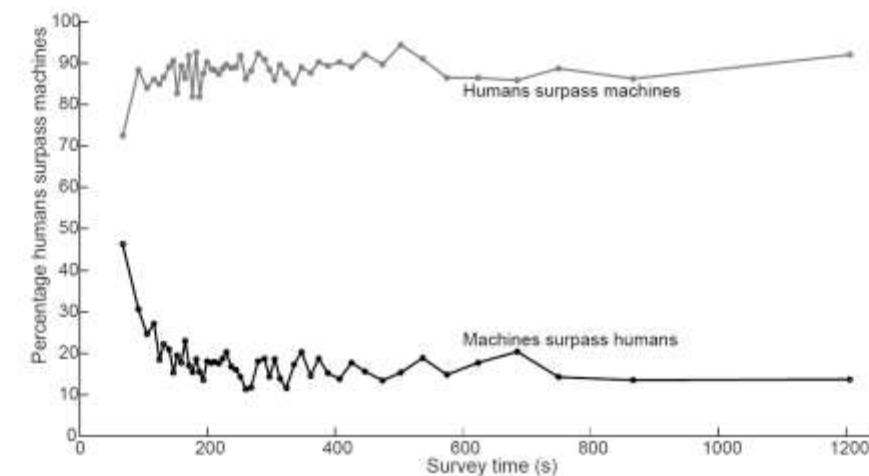


Fig. 3. The percentage of respondents who think that humans surpass machines as a function of survey time ($N = 2,941$)

- The purpose of the discussion panel is to address engineering human-AI teams from inception through sustainment to enable appropriate bi-directional trust, towards a new framework called Trust Engineering.
- Conceptual Model of Trust Engineering for Human-AI Teams - Components
 - Security
 - Adaptability
 - Communication
 - Explainability
 - Training/Knowledge
 - Assessment

Trust Engineering for Human-AI Teams

CHAIR

Neta Ezer, Ph.D., Chief Technologist for Human-Machine Teaming, Northrop Grumman Corporation

PANELISTS

Sylvain Bruni, Principal Human Systems Engineer, Aptima, Inc.

Yang Cai, Ph.D., Director of Visual Intelligence Studio, Carnegie Mellon University

Sam J. Hepenstal, Data Scientist, Defence Science and Technology Laboratory

Christopher A. Miller, Ph.D., Chief Scientist, Smart Information Flow Technologies

Dylan D. Schmorow, Ph.D., Vice President and Chief Scientist, Soar Technology, Inc.

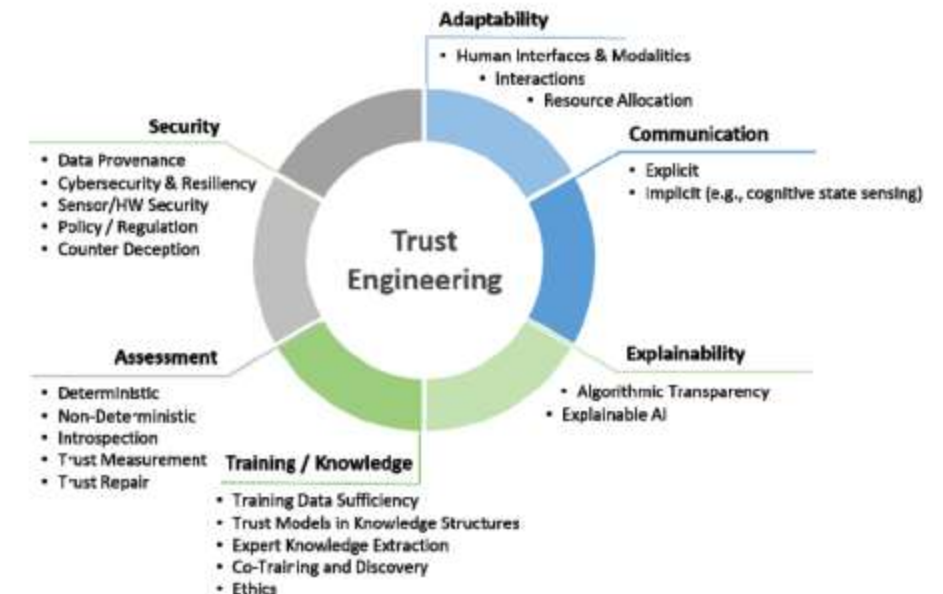


Figure 1. Conceptual Model of Trust Engineering for Human-AI Teams

© Academy of Management Annals
2020, Vol. 14, No. 2, 627–660.
<https://doi.org/10.5465/annals.2018.0057>

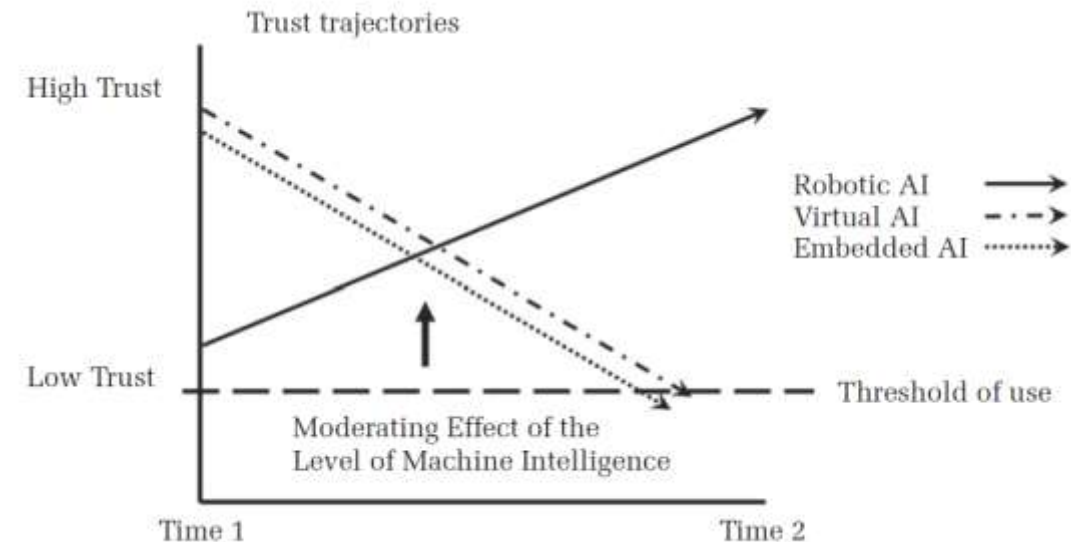
- Main Effects of Dimensions on Cognitive Trust in AI, Organized by Representation (Robotic, Virtual or Embedded AI's)
 - Tangibility
 - Transparency
 - Reliability
 - Task Characteristics
 - Immediacy Behaviors
- Main Effects of Dimensions on Emotional Trust in AI, Organized by Representation (Robotic, Virtual or Embedded AI's)
 - Tangibility
 - Anthropomorphism
 - Immediacy behaviors
- It could be useful to consider the different cognitive and emotional implications of human physical presence, conceptualizing presence as a continuum rather than as a category (i.e., colocated vs. distributed) and further exploring the moderating factors in the tangibility–trust relationship.

HUMAN TRUST IN ARTIFICIAL INTELLIGENCE: REVIEW OF EMPIRICAL RESEARCH

ELLA GLIKSON¹
Bar Ilan University

ANITA WILLIAMS WOOLLEY
Carnegie Mellon University

Artificial intelligence (AI) characterizes a new generation of technologies capable of interacting with the environment and aiming to simulate human intelligence. The suc-



- unlike people, machines do not become tired or bored [40, 98], and can take into account orders of magnitude more factors than people can
- like people, algorithms are vulnerable to biases that render their decisions “unfair”
- In this survey we identify two potential sources of unfairness in machine learning outcomes—those arising from biases in the data and those arising from the algorithms
- Therefore, it is important for researchers and engineers to be concerned about the downstream applications and their potential harmful effects when modeling an algorithm or a system
 - *Accounting for the social/political/moral implications of algorithms is a human function*
- In order to be able to address the bias issue in these application, it is important for us to know where these biases are coming from and what we can do to prevent them.

A Survey on Bias and Fairness in Machine Learning

NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI

With the widespread use of AI systems and applications in our everyday lives, it is important to take fairness issues into consideration while designing and engineering these types of systems. Such systems can be used in many sensitive environments to make important and life-changing decisions; thus, it is crucial to ensure that the decisions do not reflect discriminatory behavior toward certain groups or populations. We have recently seen work in machine learning, natural language processing, and deep learning that addresses such challenges in different subdomains. With the commercialization of these systems, researchers are becoming aware of the biases that these applications can contain and have attempted to address them. In this survey we investigated different real-world applications that have shown biases in various ways, and we listed different sources of biases that can affect AI applications. We then created a taxonomy for fairness definitions that machine learning researchers have defined in order to avoid the existing bias in AI systems. In addition to that, we examined different domains and subdomains in AI showing what researchers have observed with regard to unfair outcomes in the state-of-the-art methods and how they have tried to address them. There are still many future directions and solutions that can be taken to mitigate the problem of bias in AI systems. We are hoping that this survey will motivate researchers to tackle these issues in the near future by observing existing work in their respective fields.

CCS Concepts: • Computing methodologies → Artificial intelligence; Philosophical/theoretical foundations of artificial intelligence;

Additional Key Words and Phrases: Fairness and Bias in Artificial Intelligence, Machine Learning, Deep Learning, Natural Language Processing, Representation Learning

1 INTRODUCTION

Machine algorithms have penetrated every aspect of our daily life. Algorithms make movie recommendations, suggest products to buy, and are increasingly used in high-stakes decisions in loan applications [93], dating and hiring [18, 35]. There are clear benefits to algorithmic decision-making; unlike people, machines do not become tired or bored [40, 98], and can take into account orders of magnitude more factors than people can. However, like people, algorithms are vulnerable to biases that render their decisions “unfair” [5, 100]. In the context of decision-making, fairness is the *absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics*. Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people. A canonical example comes from a tool used by courts in the United States to make parole decisions. The software, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), measures the risk of a person to recommit another crime. Judges use COMPAS to decide whether to release an offender, or to keep him or her in prison. An investigation into the software found a bias against African-Americans:¹ COMPAS is more likely to assign a higher risk score to African-American offenders than to Caucasians with the same profile. Similar findings

¹<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

“The Need for Biases in Learning Generalizations.”

Rutgers University, New Brunswick, NJ.

http://dml.cs.byu.edu/~cgc/docs/ml_dm_tools/Reading/Need%20for%20Bias.pdf

The Need for Biases in Learning Generalizations

Tom M. Mitchell

1. Introduction

Learning involves the ability to generalize from past experience in order to deal with new situations that are “related to” this experience. The inductive leap needed to deal with new situations seems to be possible only under certain biases for choosing one generalization of the situation over another. This paper defines precisely the notion of bias in generalization problems, then shows that biases are necessary for the inductive leap. Classes of justifiable biases are considered, and the relationship between bias and domain-independence is considered.

We restrict the scope of this discussion to the problem of generalizing from training instances, defined as follows:

The Generalization Problem
Given:

1. Language of instances.
2. Language of generalizations.
3. Matching predicate for matching generalizations to instances.
4. Sets of positive and negative training instances.

Determine:

⇒ Generalization(s) consistent with the training instances.

As a concrete example of the above generalization problem, consider the task addressed by Winston’s program for learning classes of block structures (Winston 1975). Here, the language of instances is the representation used to describe example block structures. The language of generalizations is the language in which learned concepts (e.g., arch, tower) are described. The matching predicate specifies whether a given generalization applies to a given instance (e.g., whether the inferred description of an arch is satisfied by a specific block structure).

This paper addresses a deep difficulty with the generalization problem as defined above: If consistency with the training instances is taken as the sole determiner of appropriate generalizations, then a program can never make the inductive leap necessary to classify instances beyond those it has observed. Only if the program has other sources of information, or biases for choosing one generalization over the

Converted to electronic version by: Roby Joehanes, Kansas State University

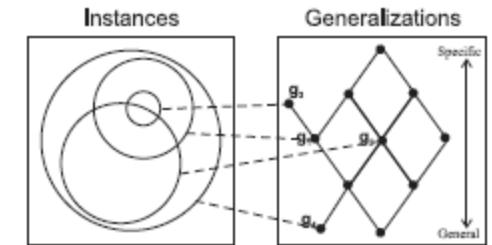


Figure 1: Relationships among Instances and Generalizations (This figure was missing from the original publication and added in 1990.)

other, can it non-arbitrarily classify instances beyond those in the training set.

In this paper, we use the term *bias* to refer to *any basis for choosing one generalization over another, other than strict consistency with the observed training instances.*

2. What is an UNbiased Generalizer

If generalization is the problem of guessing the class of instances to which the positive training instances belong, then an unbiased generalizer is one that makes no a priori assumptions about which classes of instances are most likely, but bases all its choices on the observed data. Two common sources of bias in existing learning systems are (1) the generalization language is not capable of expressing all possible classes of instances, and (2) the generalization procedure that searches through the space of expressible generalizations is itself biased.

2.1. An Unbiased Generalization Language

In considering bias in the generalization language, it is useful to view each generalization as denoting the set of instances that it matches. In figure 1, for example, g_1 and g_2 are two generalizations expressible in some generalization language, and each matches a different subset of the instances.

Relative to a given language of instances, an unbiased generalization language is then one which allows describing every possible subset of these instances. In short, an

- Although removing all biases from a generalization system may seem to be a desirable goal, in fact the result is nearly useless. An unbiased learning system’s ability to classify new instances is no better than if it simply stored all the training instances and performed a lookup when asked to classify a subsequent instance
- Progress toward understanding learning mechanisms depends upon understanding the sources of, and justification for, various biases
- Efforts to study machine learning must focus on the combined use prior knowledge, biases, and observation in guiding the learning process. It would be wise to make the biases and their use in controlling learning just as explicit as past research has made the observations and their use.

A Proposal for Identifying and Managing Bias in Artificial Intelligence

Reva Schwartz
Leann Down
Adam Jonas
Elham Tabassi

This draft publication is available free of charge from:
<https://doi.org/10.6028/NIST.SP.1270-draft>

- the presence of bias in automated systems can contribute to harmful outcomes and a public lack of trust. Managing bias is a critical but still insufficiently developed building block of trustworthiness
- Instead of viewing the challenge of AI bias within a given context or use case, a broader perspective can strike the problem of AI bias where it might be easiest to manage – within the design, development, and use of AI systems.
- One challenge rests on the reality that decisions about which data to use for these indices are often made based on what is available or accessible, rather than what might be most suitable - but difficult or impossible to utilize
- The reviewed literature suggests that the expansion of AI into many aspects of public life requires extending our view from a mainly technical perspective to one that considers AI within the social system it operates
- trust can improve if the public is able to interrogate systems and engage with them in a more transparent manner
- Whether unconscious or unintentional, pre-design is often where decisions are made that can inadvertently lead to harmful impact, or be employed to extremely negative societal ends. By not addressing the possibility of optimistic and potentially inflated expectations related to AI systems, risk management processes could fail to communicate and set reasonable limits related to mitigating such potential harms
- Identifying downstream impacts may take time and require the involvement of end-users, practitioners, subject matter experts, and interdisciplinary professionals from the law and social science.
- Not taking context into consideration during model selection can lead to biased results for subpopulations (for example, disparities in health care delivery).

Figure 1: A three-stage approach for managing AI bias



- To better identify and mitigate organizational factors which can contribute to bias, experts also suggest the use of algorithmic decision-making tools for **specific, well-defined use cases**, and not beyond those use cases
- One key issue is finding a configuration that enables a system to be used in a way that optimally leverages, instead of replaces, user expertise
- There are individual differences in how humans interpret AI model output. When system designers do not take these differences into consideration it can contribute to misinterpretation of that output
- [Appendix has a glossary of types of AI bias]

- this report concludes:
 - Bias is neither new nor unique to AI.
 - The goal is not zero risk but rather, identifying, understanding, measuring, managing and reducing bias.
 - Standards and guides are needed for terminology, measurement, and evaluation of bias.
 - Bias reduction techniques are needed that are flexible and can be applied across contexts, regardless of industry.
 - NIST plans to develop a framework for trustworthy and responsible AI with the participation of a broad set of stakeholders to ensure that standards and practices reflect viewpoints not traditionally included in AI development.
 - NIST will collaboratively develop additional guidance for assurance, governance, and practice improvements as well as techniques for enhancing communication among different stakeholder groups

Taxonomy and Definitions for Terms Related to Driving Automation

Systems for On-Road Motor Vehicles. J3016_202104.

https://www.sae.org/standards/content/j3016_202104/



SAE J3016™ LEVELS OF DRIVING AUTOMATION

		SAE LEVEL 0	SAE LEVEL 1	SAE LEVEL 2	SAE LEVEL 3	SAE LEVEL 4	SAE LEVEL 5
What does the human in the driver's seat have to do?		You <u>are</u> driving whenever these driver support features are engaged – even if your feet are off the pedals and you are not steering			You <u>are not</u> driving when these automated driving features are engaged – even if you are seated in “the driver’s seat”		
		You must constantly supervise these support features; you must steer, brake or accelerate as needed to maintain safety			When the feature requests, you must drive	These automated driving features will not require you to take over driving	
What do these features do?		These are driver support features			These are automated driving features		
		These features are limited to providing warnings and momentary assistance	These features provide steering OR brake/acceleration support to the driver	These features provide steering AND brake/acceleration support to the driver	These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met	This feature can drive the vehicle under all conditions	
Example Features		• automatic emergency braking • blind spot warning • lane departure warning	• lane centering OR • adaptive cruise control	• lane centering AND • adaptive cruise control at the same time	• traffic jam chauffeur	• local driverless taxi • pedals/steering wheel may or may not be installed	• same as level 4, but feature can drive everywhere in all conditions

Calibrating Human-AI Collaboration: Impact of Risk, Ambiguity and Transparency on Algorithmic Bias.

In: Holzinger A., Kieseberg P., Tjoa A., Weippl E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2020. Lecture Notes in Computer Science vol 12279. Springer, Cham. https://doi.org/10.1007/978-3-030-57321-8_24

- The right level of transparency can increase trust in an AI system, while inappropriate levels of transparency can lead to algorithmic bias
- the danger of algorithmic bias or algorithm aversion should motivate not only calibrating the transparency level of AI systems. An equally important focus of research is the calibration of humans for responsible usage of transparent AI systems.
- adding transparency does not have the same effect on users of an AI system, depending on their personality.
- Our results indicate that not only in human interaction but also in the context of assistive AI technology, these two factors, tolerance of risky and ambiguous uncertainty, play an important role
- Herman points out the dilemma in interpretability research: there is a tradeoff between explaining a model's decision faithfully and in a way that humans easily understand
- We believe that human-AI interaction is inherently associated with risky and ambiguous uncertainty
- transparency induces algorithmic bias, meaning annotators are more likely to replicate the ML prediction.
- Annotators that tended to be risk averse showed the most pronounced algorithmic bias
- Risk tolerant annotators were not as susceptible to transparency-induced algorithmic bias, the agreement of their annotations with the ML model predictions was less affected by the level of transparency
- transparency in itself does not necessarily have positive consequences on human-AI interaction
- Optimal calibration however requires an in-depth understanding of both parties and how they interact.
- transparency and the induced algorithmic bias can lead to worse annotation accuracy when the ML model prediction is uncertain or wrong--but we also found that algorithmic bias can lead to increased annotation accuracy when the model is certain or correct in its predictions.

Calibrating Human-AI Collaboration: Impact of Risk, Ambiguity and Transparency on Algorithmic Bias

Philipp Schmidt ^{*1} and Felix Biessmann ^{*1,2}

¹ Amazon Research, Berlin, Germany
{phschmid,biessman}@amazon.com

² Einstein Center Digital Future, Berlin, Germany
felix.biessmann@beuth-hochschule.de

Abstract. Transparent Machine Learning (ML) is often argued to increase trust into predictions of algorithms however the growth of new interpretability approaches is not accompanied by a growth in studies investigating how interaction of humans and Artificial Intelligence (AI) systems benefits from transparency. The right level of transparency can increase trust in an AI system, while inappropriate levels of transparency can lead to algorithmic bias. In this study we demonstrate that depending on certain personality traits, humans exhibit different susceptibilities for algorithmic bias. Our main finding is that susceptibility to algorithmic bias significantly depends on annotators' affinity to risk. These findings help to shed light on the previously underrepresented role of human personality in human-AI interaction. We believe that taking these aspects into account when building transparent AI systems can help to ensure more responsible usage of AI systems.

Keywords: transparent AI · machine learning · HCI · risk affinity

1 Introduction

Decision support by Machine Learning (ML) models has become ubiquitous in everyday life. Responsible usage of such assistive technology requires an appropriate level of trust into ML systems [7]. Trust into this technology is often argued to require interpretability of ML predictions. This is why transparent ML methods have become an active focus of research. A central research question in this field is how methodological advances translate into improvements in human-AI interaction and trust. Transparent ML research aims at better explaining ML predictions and systems to humans, but it is difficult to assess how much human-AI interaction profits from scientific advances in the field. Many studies proposing new interpretable ML methods do report results with human-in-the-loop experiments to test the impact of their proposed method on

* equal contribution

“Designing Trustworthy AI for Human-Machine Teaming.”

Presented at AAAI FSS-19: Artificial Intelligence in Government and Public Sector, Arlington, Virginia, USA

Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development

Carol J. Smith

Carnegie Mellon University's Software Engineering Institute, Pittsburgh, PA
cjsmith@sei.cmu.edu

- Diverse teams are needed to build trustworthy artificial intelligent systems, and those teams need to coalesce around a shared set of ethics
- usability testing can help determine if the audience understands how the AI system works and complies with the HMT Framework
- The HMT Framework is to be used in conjunction with the team's technology ethics code, as a tool to support the inspection process and gauge if the process and the resulting system is indeed trust-able
- There are four themes in the HMT Framework for Design-ing Ethical AI Experiences and they will be discussed in de-tail in this paper. AI systems need to be created to be:
 - 1) Accountable to humans
 - 2) Cognizant of speculative risks and benefits
 - 3) Respectful and secure
 - 4) Honest and usable
- final decisions that affect a person's life, quality of life, health, or reputation, must be made by a human being
- The diverse team needs to make time to identify the full range of harmful and malicious use of the AI system.
- AI systems need to be accountable and transparent to the humans who made them, monitor them, and use them. People should be able to verify what the AI system is doing, and why, in a timely manner
- An honest and usable system values transpar-ency with the goal of engendering trust of everyone interacting with it. This includes explaining the AI system and its limitations in language that the audience understands
- All bias needs to be recognized and explained clearly in AI systems [Are humans better at recognizing bias than machines?]

Abstract

Artificial intelligence (AI) holds great promise to empower us with knowledge and augment our effectiveness. We can—and must—ensure that we keep humans safe and in control, particularly with regard to government and public sector applications that affect broad populations. How can AI development teams harness the power of AI systems and design them to be valuable to humans?

Diverse teams are needed to build trustworthy artificial intelligent systems, and those teams need to coalesce around a shared set of ethics. There are many discussions in the AI field about ethics and trust, but there are few frameworks available for people to use as guidance when creating these systems. The Human-Machine Teaming (HMT) Framework for Designing Ethical AI Experiences described in this paper, when used with a set of technical ethics, will guide AI development teams to create AI systems that are accountable, de-risked, respectful, secure, honest, and usable.

To support the team's efforts, activities to understand people's needs and concerns will be introduced along with the themes to support the team's efforts. For example, usability testing can help determine if the audience understands how the AI system works and complies with the HMT Framework. The HMT Framework is based on reviews of existing ethical codes and best practices in human-computer interaction and software development.

Human-machine teams are strongest when human users can trust AI systems to behave as expected, safely, securely, and understandably. Using the HMT Framework to design trustworthy AI systems will provide support to teams in identifying potential issues ahead of time and making great experiences for humans.

Introduction

Diverse teams are needed to build trustworthy artificial intelligent (AI) systems, and those teams need to coalesce around a shared set of ethics. The experience of working with people who are significantly different from us can be challenging, and studies have shown (Rock, 2019) diversity is worth the effort, as it increases our capacity for innovation and creative thinking. The shared set of ethics acts as a central point from which commonalities can be drawn, strengthening the team and their work.

Having a set of shared professional ethics helps smooth differences, to create an inclusive environment for the entire organization. In addition, a set of technology ethics, focused

on the work being done, is equally important. The author recommends initially adopting an existing set of ethics from a reputable organization that is focused on computing and AI. The Montréal Declaration for a Responsible Development of Artificial Intelligence 2018: <https://www.montrealdeclaration-responsibleai.com> (Université de Montréal, 2018; Fjeld, et al, 2019; Hagendorff, 2019), is a respected starting place. As the organization matures, efforts can be made to adapt and customize a set of ethics to the needs of the team and the work.

An ethics code without usable guidance and/or training is less likely to be implemented as intended. The HMT Framework is to be used in conjunction with the team's technology ethics code, as a tool to support the inspection process and gauge if the process and the resulting system is indeed trust-able. The HMT Framework is meant to bridge the gap between ethical stances such as “do no harm,” and the reality of applying that stance to making new technology for humans to interact with.

To be effective in this work, teams must be diverse with regard to gender, race, education, thinking process, disability status, and more, as well as their skill set and problem framing approach. This includes a multi-disciplinary team mixing machine learning experts, programmers, system architects, product managers, and those who work as curiosity experts (CE), focused on understanding the situation, the constraints, and abilities of the people who will use the system and how it would be used. CEs comprise a broad range of job titles and skills including human-computer interaction and human-machine interaction professionals, cognitive psychologists, digital anthropologists, and user experience (UX) researchers. People without these job titles and skills should also participate in (and potentially lead) CE-related activities described in this paper. These CE activities will enable the team to uncover potential issues before they arise. The goal with bringing these diverse individuals together is to reduce bias in the system and to account for a broad set of unintended consequences. “Building a company or team that actively breaks down systemic prejudice for its own sake translates to building products that avoid harming large swaths of the population,” stated UX expert Dan Brown in his blog post “UX in the Age of Abusability” (<https://greenonions.com/ux-in-the-age-of-abusability-797cd01f6b13>).

Walker, Guy & Stanton, Neville & Salmon, Paul & Jenkins, Daniel. (2008).
**A Review of Sociotechnical Systems Theory: A Classic Concept for
New Command and Control Paradigms.**
Theoretical Issues in Ergonomics Science. 9. 479-499.
10.1080/14639220701635470.

- the interaction of social and technical factors creates the conditions for successful (or unsuccessful) system performance
- optimisation of either socio, or far more commonly the technical, tends to increase not only the quantity of unpredictable, 'un-designed', nonlinear relationships, but those relationships that are actually injurious to the system's performance
- despite attempts to impose deterministic behaviour on an organisation and its environment, the resultant interlinked entities, or 'system', is actually very hard to maintain in a fixed state (or as a 'closed system')

**A Review of Sociotechnical Systems Theory: A Classic Concept
for New Command and Control Paradigms**

Dr Guy H. Walker, Prof. Neville A. Stanton, Paul M. Salmon & Daniel P. Jenkins.

Human Factors Integration Defence Technology Centre (HFIDTC)

Brunel University, BIT Lab, Uxbridge, [UK]UB8 3PH

ABSTRACT

Command and control is the management infrastructure for any large, complex, dynamic resource system (Harris & White, 1987). Traditional military command and control is increasingly challenged by a host of modern problems, namely environmental complexity, dynamism, new technology, and competition that is able to exploit the weaknesses of an organisational paradigm that has been dominant since the industrial revolution. The conceptual response to these challenges is a new type of command and control organisation called Network Enabled Capability (NEC). Although developed independently, NEC exhibits a high degree of overlap with concepts derived from sociotechnical systems theory, a fact that this paper aims to explore more fully. Uniquely, what sociotechnical theory brings to NEC research is a successful fifty year legacy in the application of open systems principles to commercial organisations. This track record is something that NEC research currently lacks. The paper reviews the twin concepts of NEC and sociotechnical systems theory, the underlying motivation behind the adoption of open systems thinking, a review of classic sociotechnical studies and the current state of the art. It is argued that 'classic' sociotechnical systems theory has much to offer 'new'

Government use and distribution.

“Reflections on the 1951 Fitts list: Do humans believe now that machines surpass them?” 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences.

- present-day machines are considered to surpass humans in respect to detection, perception, and long-term memory while Fitts argued that the opposite held true.
- As regards to detection (Statement 1 in Table 1), sensors have become more precise for all base quantities
- Computational power (Statement 10) and the ability to store very large amounts of information (Statement 4) follow an exponential temporal trend in accord with Moore’s law
- A total of 249 engineering MSc students were asked to indicate whether humans surpass machines or whether machines surpass humans, for each of the 11 statements of the Fitts list, using an electronic voting system. 2,948 respondents from 103 countries did the same via a crowdsourcing facility. [\[Could we use an analogous methodology for an AI-relevant Fitts list?\]](#)

6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015

Reflections on the 1951 Fitts list: Do humans believe now that machines surpass them?

J.C.F. de Winter^{a,*}, P.A. Hancock^b

^aDepartment Biomechanical Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands
^bMIT¹ Laboratories, Department of Psychology and the Institute for Simulation and Training, University of Central Florida, Orlando, FL USA

Abstract

In 1951, Paul Fitts “surveyed the kinds of things men can do better than present-day machines, and vice versa” and introduced a list of statements, now commonly known as the Fitts list. A criticism sometimes raised with respect to the Fitts list is that it is outdated. Hence, it is important to address the issue of modernity, especially in light of the increasing rate of technological advance. A total of 249 engineering MSc students were asked to indicate whether humans surpass machines or whether machines surpass humans, for each of the 11 statements of the Fitts list, using an electronic voting system. 2,948 respondents from 103 countries did the same via a crowdsourcing facility. The results indicated that present-day machines are considered to surpass humans in respect to detection, perception, and long-term memory while Fitts argued that the opposite held true. However, a closer reading of the 1951 report reveals that some of the supposed disagreements were already anticipated at that time.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).
Peer-review under responsibility of AHFE Conference

Keywords: Function allocation; crowdsourcing; CrowdFlower; clickers; electronic voting system

1. Introduction

Function allocation is perhaps the most critical issue in all of Human Factors research. The question whether tasks should be performed by humans or by machines has been exercising the minds of those researching human

* Corresponding author. Tel.: +31-15-2786794.
E-mail address: j.c.f.devries@tudelft.nl

Initial validation of the trust of automated systems test (TOAST)

Heather M. Wojton, Daniel Porter, Stephanie T. Lane, Chad Bieber, and Poornima Madhavan

Institute for Defense Analyses

ABSTRACT

Trust is a key determinant of whether people rely on automated systems in the military and the public. However, there is currently no standard for measuring trust in automated systems. In the present studies, we propose a scale to measure trust in automated systems that is grounded in current research and theory on trust formation, which we refer to as the Trust in Automated Systems Test (TOAST). We evaluated both the reliability of the scale structure and criterion validity using independent, military-affiliated and civilian samples. In both studies we found that the TOAST exhibited a two-factor structure, measuring system understanding and performance (respectively), and that factor scores significantly predicted scores on theoretically related constructs demonstrating clear criterion validity. We discuss the implications of our findings for advancing the empirical literature and in improving interface design.

ARTICLE HISTORY

Received 30 July 2019
Accepted 19 March 2020

KEYWORDS

Trust in Automation;
Human-Machine Trust; Trust
Scale; Confirmatory Factor
Analysis

Systems that act with little to no input from humans are increasingly common. The U.S. military, for example, uses automated systems to perform search and rescue missions and to assume control of aircraft to avoid ground collision. The public uses these systems widely to perform tasks as trivial as vacuuming to those as critical as regulating nuclear power plants. Automated systems use sensors and computer programming to perform tasks with little or no human intervention (Boulain & Verbruggen, 2017). Ideally, these systems should improve people's safety and performance by completing complex, repetitive tasks efficiently and aiding or replacing humans in hazardous environments. In reality, however, automated systems are imperfect and can behave in ways that produce unintended consequences. This imperfection can result in misuse or disuse of the systems under certain conditions.

Trust is a key determinant of whether people will rely on automated systems (Hoff & Bashir, 2015; Lyons & Guznov, 2019; Lyons et al., 2016). Research demonstrates that people rely on automation when they believe that it will help them achieve their goals in situations characterized by uncertainty and vulnerability (Lee & See, 2004). Accidents can occur, however, when people trust automated systems inappropriately. The Costa Concordia cruise ship sunk in 2012, for example, because the captain placed too little trust in the ship's automated navigation system and elected not to rely on it. He manually diverged from the route it selected, crashing into a shallow reef and killing 32 passengers (Levs, 2012). It is critical, therefore, that automated systems are designed to facilitate appropriate levels of trust.

Currently, there is no standard for measuring trust in automation. In general, researchers measure it using custom scales or validated measures of theoretically related constructs, making it difficult to draw comparisons across studies or craft criteria to determine if automation is facilitating appropriate levels of trust (Merritt, 2011). Furthermore, the published scales that do exist perform inconsistently, exhibiting unreliable scale structure (Jian et al., 2000). In the present studies, we

Table 2. Revised multi-item scale.

System Understanding	System Performance
• I understand what the system should do.	• The system helps me achieve my goals.
• I understand the limitations of the system.	• The system performs consistently.
• I understand the capabilities of the system.	• The system performs the way it should.
• I understand how the system executes tasks.	• I am rarely surprised by how the system responds.
	• I feel comfortable relying on the information provided by the system.

1. Don't expect Machine learning to figure out what problems to solve
2. Ask yourself if ML will address the problem in a unique way [knowing what humans are good at and what AI is good at will help here]
3. Fake it with personal examples and wizards
4. Weigh the costs of false positives and false negatives
5. Plan for co-learning and adaptation
6. Teach your algorithm using the right labels
7. Extend your UX family, ML is a creative process

AI will be limited in its ability to act alone for the foreseeable future

- Perceptual limitations
- Brittleness
- Hidden biases
- No model of causation

- AI will need to act effectively with humans, but there are many known problems
 - Automation confusion
 - Poor SA & out-of-the-loop
 - Irony of automation
 - Human decision biasing
 - Skills degradation

Need for team coordination increases as capabilities of technology increase

Human-Centered AI

- **Designing an AI system to work well as a teammate increases human-centeredness**
- **Augments human capabilities and raises performance beyond that of either entity**

Need Improved Models for Human-AI Team Processes

- **Team heterogeneity**
 - Need for long-term, distributed, agile teams that exhibit function allocation, interdependency management, and responsibility overlap
- **Shared cognition**
 - Shared mental models provides an understanding of teammates' capabilities, limitations, current goals and needs, and current and future performance.
- **Communication and Coordination**
 - Both explicit and implicit, anticipation of information needs & tasks of teammates, requires a deep model of teammates

- **Goal Alignment**
 - Desired goal state actions need to support
 - Requires active goal switching based on prioritization
- **Function Allocation/Re-allocation**
 - Assignment of functions and tasks across team
 - Dynamic reassignment based on capabilities, status
- **Decision Communication**
 - Selection of strategies, plans and actions needed to bring world into alignment with goals
- **Task Alignment**
 - Coordination of inter-related tasks for effective overall operations

Shared Situation Awareness

es, and intentions of fellow teammates

- Research Objective 7-3: Cooperation as a Measure of Trust. In teaming structures, behavioral measures of cooperation would be useful to employ, to understand when trust matters beyond reliance and compliance in function allocation structures. Team science literature provides a useful framework for understanding human-AI teaming (i.e., interactions with increasingly capable automation) because teams in general tend to be comprised of members with high levels of autonomy, meaning team members do not have complete control over other team members (NRC, 2015). Therefore, an assumption of teaming is that it is not deterministic (i.e., not akin to choreographing a collection of autonomous agents), but that coordination emerges from the team's interactions (Cooke et al., 2013) with trust playing a central role (i.e., in teamwork). In the function allocation approach outlined in early conceptualizations of human-automation systems (Fitts, 1951; Sheridan and Verplank, 1978), which remains prevalent in modern applications of AI with well-defined roles narrowly scoped to a particular task, reliance and compliance may make sense as primary behavioral outcomes of interest when it comes to human trust. However, coordination to achieve a specific goal in laboratory studies of teaming needs to be distinguished from coordination in more complex environments, in the field or in simulated settings, that involve uncertainty and changing goals. These environments often demand a dynamic process of trust that is central to coordination.



Research Objective 6-1: Human-AI Team Task Sharing. Research is needed to determine improved methods for supporting collaboration between humans and AI systems in shared functions, at intermediate levels of automation. It would be beneficial for this research to focus on a more detailed understanding of the ways that people and AI systems can share tasks (Miller, 2018), and to explore various methods for combining humans and AI systems, to improve cognitive performance and resilience to errors and unforeseen conditions (Cummings et al., 2011; Smith, 2017, 2018).



AI is not the solution to every problem (though if you ask an AI engineer, you’ll get a different answer—ML is their hammer, so everything is a nail). By understanding what humans and AI are good at (and by applying good UX and Design Thinking principles), we can ensure that the user and the problem remain the focus, not just the AI technology

Why use design thinking?

- Ensures the **user** and **problem** are the focus of the project, not the technology
- Aligns all project team members on a common direction
- Validates whether we even need to apply AI to solve the problem

Human vs. AI prediction

Human prediction	AI prediction
Limited memory and processing ability	Ability to process vast amounts of data
Decisions influenced by many factors – emotions, biases, physical state	Consistent decisions not influenced by external factors
Lean on mental heuristics to reach decisions that fit with our concept of the world	Makes decisions based solely on data (which can contain biases)
Can apply commonsense reasoning to novel situations	Cannot reason beyond the data it was trained on

Automation vs. augmentation

- AI can easily learn routine tasks which follow consistent patterns
- Other tasks require understanding of context, or ability to reason
- Use cases for AI fall into two categories:
 - **Automation** – replacing humans
 - **Augmentation** – supporting humans

Communicating performance

Outcome Metrics

- Refers to the desired business impact for your users
- Stated in terms of the expected impact (which is often \$)

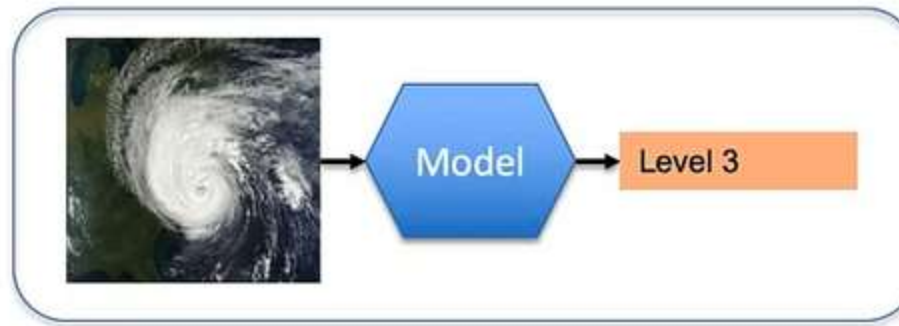
Output Metrics

- Refers to the desired output from the model
- Measured in terms of a model performance metric

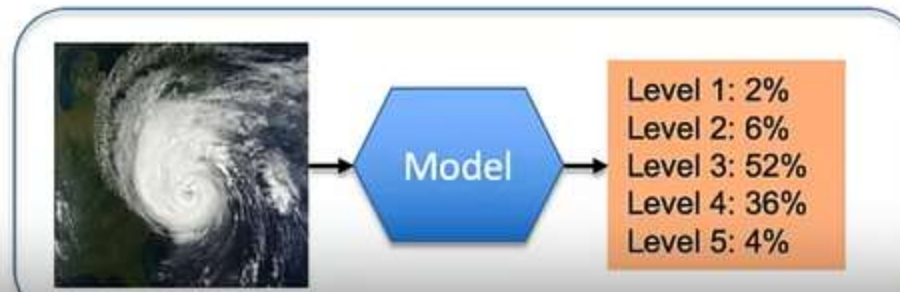
How to use metrics to inspire trust:

- 1) Communicate outcomes for previous users
- 2) Track and communicate model metrics for current user

Presenting confidence



versus



Human-in-the-loop

- Human quality control of model outputs can flag issues before they impact users
- Particularly important early in model rollout
- Opportunity to correct model or provide alternative support to customer
- Should use carefully – not try to guess when model is right or wrong

Wrap-Up

- AI learns and makes predictions very differently from humans
- Augmentation approaches such as triage and decision support can take advantage of complimentary skillsets
- Intentional focus on building model trust and proper onboarding can ensure adoption

Cognitive collaboration: Why humans and computers think better together.

Deloitte Insights <https://www2.deloitte.com/us/en/insights/deloitte-review/issue-20/augmented-intelligence-human-computer-collaboration.html>

- With intelligence augmentation, the ultimate goal is not building machines that think like humans, but designing machines that help humans think better.
- Algorithms are reliable only to the extent that the data used to train them are sufficiently complete and representative of the environment in which they are to be deployed. When this condition is not met, all bets are off.
- When routine tasks can be encoded in big data, it is a safe bet that algorithms can be built to perform them better than humans can. But such algorithms will lack the conceptual understanding and commonsense reasoning needed to evaluate novel situations. They can make inferences from structured hypotheses but lack the intuition to prioritize which hypothesis to test in the first place.
- When the cases are frequent, unambiguous, and similar across time and context—and if the downside costs of a false prediction are acceptable—algorithms can presumably automate the decision. On the other hand, when the cases are more complex, novel, exceptional, or ambiguous—in other words, not fully represented by historical cases in the available data—human-computer collaboration is a more plausible and desirable goal than complete automation.
- “Marvin was advocating what’s called “commonsense reasoning.” Machines have shown essentially no examples of doing that. Therefore, they are complements to people. People are actually not so bad at that. However, they are somewhat lousy at tuning things and keeping exact accounts of stuff. Machines are good at that.” – Sandy Pentland (interview excerpt)