An Investigation of Human-Model Interaction for Model-Centric Decision-Making

by

Erling Shane German

B.S. Systems Engineering
United States Air Force Academy, 2015

Submitted to the Institute for Data, Systems, and Society
in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Technology and Policy

at the

Massachusetts Institute of Technology

June 2017

Signature of Author…….
Technology and Policy Program
May 12, 2017

Certified by…….
Donna H. Rhodes
Principal Research Scientist, Sociotechnical Systems Research Center
Director, Systems Engineering Advancement Research Initiative
Thesis Supervisor

Accepted by…….....
Munther Dahleh
William A. Coolidge Professor, Electrical Engineering and Computer Science
Director, Institute for Data, Systems, and Society

An Investigation of Human-Model Interaction for Model-Centric Decision-Making

by
E. Shane German

Submitted to the Institute for Data, Systems, and Society on May 12, 2017
in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Technology and Policy

Abstract

This thesis presents an investigation of human-model interaction in relation to model-centric decision-making. Models are abstractions, or simplifications, of reality that humans use to augment their ability to make sense of the world, anticipate future outcomes, and make decisions. This thesis focuses on models that aid decision-making in the design and operation of technological systems. Model-centric engineering is transforming traditional engineering towards a paradigm of comprehensive, integrated model use throughout the lifecycle of complex systems. This model-centric shift aims to increase the efficiency and efficacy of system decision-making. Without appropriately considering and designing for the human element, however, model-centric engineering will fail to achieve its desired results. Enabling effective human-model interaction, therefore, is crucial for realizing the value that models and model-centric engineering practice can provide. Advances in model technology and computational resources have been steadily made, however, the many facets of the human-model interaction experience remain relatively unexplored.

Through empirical and qualitative methods, this thesis presents an exploration of human-model interaction in an effort to identify decision-making challenges, and appropriate mitigations, for individuals in model-centric environments. Learning from existing literature and past situations with similar considerations is a useful place to start in investigating the human aspects. Two analogy case studies reveal relevant individual and organizational challenges that may affect human-model interaction and decision-making within model-centric environments. An expert interview-based study yields empirical insight from thirty experts into sociotechnical factors that influence the trust and use of models by various types of actors within the model-centric decision-making process. Additionally, as automation, autonomy, and artificial intelligence (AI) will likely play key roles in successful model-centric engineering, relevant literature-based considerations are presented for how the capabilities of AI and autonomy may relate to a model-centric context. This cumulative research is ultimately distilled into twenty-nine descriptive and prescriptive heuristics for enabling effective human-model interaction and model-centric decision-making. These heuristics emerged from the voice of the experts interviewed, as well as from case studies and literature analyzed.

Policy considerations based on this investigation are discussed, along with a suggested strategy of planned adaption for model-centric policymaking. Overall, this research aims to generate grounded theory to motivate and guide future research and development for enabling effective human-model interaction and model-centric decision-making.

Thesis Supervisor: Dr. Donna H. Rhodes
Title: Principal Research Scientist, Sociotechnical Systems Research Center

Disclaimer

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense through the Systems Engineering Research Center (SERC) under Contract HQ0034-13-D-0004. SERC is a federally funded University Affiliated Research Center managed by Stevens Institute of Technology. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Department of Defense.

In memory of my Dad

Time flies like an arrow; fruit flies like a banana

-Groucho Marx

Acknowledgements

This thesis is not just a culmination of nearly two years of research, but perhaps even more so of the relationships that shaped my experience at MIT. Whether you call it blind luck, happy coincidence, or divine providence, I cannot fully express how thankful I am to have ended up under the mentorship of my advisor, Dr. Donna Rhodes. You taught me how to research, and perhaps even more importantly, that I am capable of research. Regardless of the impact of this work, I can assure you that it is a testament to the impact that your mentorship and guidance have had upon me. I also want to thank the other fellow members of SEAri that I have interacted with throughout my time here – your feedback, ideas, and comradery have significantly shaped my thoughts and research. Specifically, I would like to thank Paul La Tour for the support, insight, and hours of enjoyable conversation that you provided throughout my first year. In regards to this final thesis draft, I need to thank my oldest sister, Sara, for the hours of editing you so willing offered. Despite being a med student who is far busier than I, your editing prowess vastly improved the quality of this thesis.

I also want thank my friends and mentors throughout the Boston area – while the research and classwork were important and (mostly) enjoyable, these relationships added so much more purpose and meaning to my time here. Specifically, to Dave, Jordan, and Meg: I owe you all a debt of gratitude – my grad school experience would not have been the same without you. To Mom, Sara, Jilanne, Kevin, and Cindy: I love you all so much, thank you for the love and support you add to my life – I would not be the person I am without you. To my earthly father in his new home: thank you for instilling in me the integrity and work ethic that guides my life. I love and miss you, Dad. To my Heavenly Father: thank you for the abilities, opportunities, love, and salvation you have so undeservedly given me. I look forward to exploring the next chapter of this story of life that you have written for me.

Biographical Note

Erling Shane German earned a Bachelor of Science degree in Systems Engineering from the U.S. Air Force Academy in 2015. Following graduation from the Academy, Shane began graduate studies in the Technology and Policy Program at MIT, during which time he worked as a research assistant in the Systems Engineering Advancement Research Initiative. This thesis marks the culminating work of his research and studies at MIT.

Born and raised in Imperial, Nebraska, Shane is a son of Scot and Linda German, and a brother to four siblings. Shane grew up working on the family farm and ranch – during which time his favorite activities included not working with cattle, and if required to do so, flying with his dad in the airplane they used to move the cattle. He was homeschooled until his sophomore year of high school, at which time he attended, and later graduated from, Chase County High School. Deciding he wanted to fly planes in the Air Force, Shane applied to, and later attended, the Air Force Academy. During his time as an Academy cadet, Shane was a member of the U.S. Air Force Parachute Team, the Wings of Blue. Upon graduating from MIT, Shane will continue pursuing his dream of becoming an Air Force pilot by attending pilot training in Wichita Falls, Texas.

Shane's life is defined by relationships: first, with his Lord and Savior Jesus Christ, and second, with his family and friends. While his priorities in life get misplaced far too often, Shane desires to continually grow as a person so that he can better love God and those he interacts with throughout his life.

Table of Contents

List of Figures

List of Tables

1 Introduction

This is a thesis about humans, models, and the interactions between the two. Models are abstractions, or simplifications, of reality that humans use to augment their ability to make sense of the world, anticipate future outcomes, and make decisions. This thesis focuses on models that aid decision-making in the design and operation of technological systems. Model-centric engineering is transforming traditional engineering towards a paradigm of comprehensive, integrated model use throughout the lifecycle of complex systems. This model-centric shift aims to increase the efficiency and efficacy of system decision-making. Without appropriately considering and designing for the human element, however, model-centric engineering will fail to achieve its desired results. Enabling effective human-model interaction, therefore, is crucial for realizing the value that models and model-centric engineering practice can provide. Advances in model technology and computational resources have been steadily made, however, the many facets of the human-model interaction experience remain relatively unexplored. Through empirical and qualitative methods, this thesis presents an investigation of human-model interaction in an effort to identify decision-making challenges, and appropriate mitigations, for individuals in model-centric environments. Ultimately, guiding heuristics are presented to inform effective model-centric practice and policy.

1.1 4 C's Conceptual Model

Figure 1 presents a conceptual model for framing the discussion of model-centric research and development. The "4 C's" within the model include capabilities, competencies, cautions, and controls. "Capabilities" represent desired outputs of model-centric engineering. In other words, capabilities are what a shift to a model-centric paradigm hopes to accomplish; they are the envisioned outcomes. "Competencies" refer to the knowledge, skills, abilities, and enabling technologies needed to achieve these capabilities. Competent individuals are required both to develop the necessary enabling technologies and to use them appropriately. A certain set of competencies are required before the desired capabilities can be achieved; in essence, the capabilities desired drive the competencies required. As with any transformation, however, challenges exist that hinder frictionless achievement of desired capabilities and needed competencies.

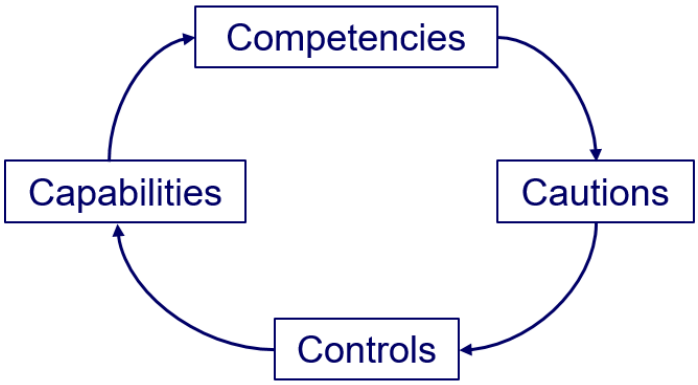


Figure 1. The 4 C's of Model-Centric Engineering Research and Development

"Cautions" consider these challenges in an effort to prevent viewing capabilities and competencies in a naïve manner. This thesis specifically investigates challenges for effective human-model interaction that could hinder effective model-centric decision-making. Cautions ultimately feed into "controls" – policies and decisions that proactively address potential challenges and develop a more nuanced and effective vision for the capabilities and competencies. Policies offer a means for structuring an environment to achieve

desired outcomes. Accordingly, those in "control" of organizational policies are in an influential position to not only develop the vision and goals for model-centric engineering, but to do so in an informed and effective manner. This thesis primarily focuses on exploring and developing relevant cautions for human-model interaction that aim to inform more effective controls for model-centric engineering.

1.2 Motivation: Model-Centric Engineering

The research discussed in this thesis explores various dimensions of enabling effective model-informed decisions, as motivated by the increasing need for individuals and teams to make decisions based on models and model-generated information. Models are abstractions of reality than can come in a variety of forms and formats, but fundamentally are encapsulations of reality that humans use to augment their ability to make sense of the world, anticipate future outcomes, and make decisions. As systems become more complex, increasingly innovative ways are needed to capture and abstract this complexity into manageable, meaningful information. Capturing complexity ironically implies increasing the complexity of models, which creates its own challenges. Among many others, these challenges include reasoning, comprehension and collaborative decision-making in the face of uncertainty, combining artificial (model-generated) and real data, and effectively utilizing vast amounts of information.

Models are increasingly used to drive major acquisition and design decisions, yet model developers, analysts, architects, program managers and senior decision-makers are faced with various challenges. Blackburn, Cloutier, Witus, Hole, and Bone (2015) captures many of these challenges in an investigation of the technical feasibility of radically transforming systems engineering through model-centric engineering (Blackburn, Bone, Witus, 2015). Digitized legacy systems and new digital system models will provide the basis for designing and evolving systems in the future (West and Pyster, 2015). This reinforces the criticality of models as assets and necessitates change in model-related policy and practices (Zimmerman, 2015a). The Model-Centric Engineering Forum conducted by the US Department of Defense (DoD) Systems Engineering Research Center (SERC) in May 2016 fostered a dialogue between industry, government, and academia on the current state of practice and vision for transformation (Clifford, Blackburn, Verma, Zimmerman, 2016).

The Interactive Model-Centric Systems Engineering (IMCSE) research program, initiated in 2014 under the sponsorship of the DoD SERC, aims to inform and contribute methods, processes and tools to improve human-model interaction in support of accelerating the transition of systems engineering to a more model-centric discipline (Rhodes and Ross, 2015; Rhodes and Ross, 2016a; Rhodes et al., 2017). IMCSE advances knowledge relevant to human interaction with models and model-generated information, drawing from relevant knowledge from other fields (e.g., cognitive science, visual analytics, data science) and placing it within the context of systems engineering. Additionally, IMCSE aims to generate knowledge that will impact human effectiveness in model-centric environments of the future (Rhodes and Ross, 2016b). As part of this exploration into human-model interaction, German and Rhodes (2016) examines the transition from traditional aircraft cockpits to modern glass cockpits as an analogy case, indicating information abstraction and automation led to new cognitive and perceptual challenges.

Significant progress continues to be made in the theory and practice of model-based engineering, yet little attention has been given to the complexities of human-model interaction. An open area of inquiry relates to the various facets of humans interacting with models and model-generated information throughout the lifecycle. The 2015 IMCSE Pathfinder Workshop (Rhodes and Ross, 2015) seeded a research agenda around the topic of human-model interaction, identifying research needs from both a model-centric (technology) perspective and an interaction (human) perspective. Participants agreed that progress has been made on standards, methods and techniques for model-based systems engineering, yet little attention has

been given to human-model interaction. A science of human-systems interaction has emerged, but its focus is on operational systems. The 2015 IMCSE Pathfinder Workshop validated the belief that improving human-model interaction would significantly improve model-centric engineering (Rhodes and Ross, 2015). Additionally, a 2016 workshop report sponsored by the National Science Foundation (NSF), the National Aeronautics and Space Administration (NASA), the Air Force Office of Scientific Research (AFOSR), and the National Modeling and Simulation Coalition (NMSC), highlights the need for understanding the individuals involved in the modeling process and how these individuals affect model development and usage. Central to this topic is the need to understand what engenders trust in models. While anecdotal stories of success and failure exist, empirical studies are needed to truly understand the many facets of human decision-making in model-centric engineering.

Additional motivation for exploring human-model interaction stems from the DoD's interest in artificial intelligence (AI), autonomous capabilities, and promoting effective human-machine interaction. Recent initiatives within the DoD seek to leverage advances in technology, specifically in artificial intelligence and autonomous capabilities, to maintain strategic and tactical advantages over potential adversaries. Deputy Secretary of Defense Robert Work points to futures of "human-machine collaboration" and "human-machine teaming" where collaboration with advanced technology will "help a human make better decisions" (Pomerleau, 2016). Work understands that such a shift requires time and effort, which includes examining "what we can do, how we train our people, how our people react" (Pomerleau, 2016). The vision of model-centric engineering shares similar goals by seeking to leverage technology to develop more effective models and modeling environments for collaborative use by decision-makers. The possibilities for what can be modeled are endless: model-centric decision-making collaboration could potentially span areas from strategic acquisition decision-making to wartime strategic, operational, and tactical decision-making. Research and development in human-model interaction and collaboration has potential to contribute to elements of the DoD's technological strategy.

1.2.1 Desired Capabilities

According to Blackburn et al. (2017), model-centric engineering can be thought of as an "overarching digital engineering approach that integrates different model types with simulations, surrogates, systems and components at different levels of abstraction and fidelity across disciplines throughout the lifecycle." In short, it represents a transformative shift that aims to "do everything with models," and more specifically, digital models (Blackburn, Cloutier, Witus, and Hole, 2014). In an engineering world traditionally formed around a document-centric acquisition process, model-centric engineering envisions shifting towards a "dynamic digital model-centric ecosystem" (Zimmerman, 2015b). Proponents of this ecosystem imagine digital models integrated together to serve as the "single source of truth" for all engineering decisions within the lifecycle (Zimmerman, 2015b).

A recent DoD Government-Industry forum (Clifford et al., 2016) cites four areas of benefit provided by a shift to model-centric practices:

1. Improved Acquisition. Accepting digital deliverables could improve the government's understanding of a project's status and risk along with allowing it to "validate" the contractor's deliverables.
2. Improved Efficiency and Effectiveness. Using a digital system model could reduce time and effort in the performance of existing tasks.
3. Improved Communication; Better Trade-Space Exploration; Reduced Risk. Using ontology-based information models to translate and extract useful information between a variety of models and model types could allow for improved communication among specialists.

4. Improved Designs and resulting Systems and Solutions. Being able to understand the impact of requirement and/or design decisions early on could help improve overall system design and identify adverse consequences before committing to a design choice.

Digital thread and digital system models are key visionary products that represent desired capabilities offered through model-centric engineering:

> Digital Thread: An extensible, configurable and component enterprise-level analytical framework that seamlessly expedites the controlled interplay of software, authoritative data, information, and knowledge in the enterprise data-information-knowledge systems, based on the Digital System Model template, to inform decision makers throughout a system's life cycle by providing the capability to access, integrate and transform disparate data into actionable information." (Zimmerman, 2015c)

> Digital System Model: A digital representation of a defense system, generated by all stakeholders, that integrates the authoritative data, information, algorithms, and systems engineering processes which define all aspects of the system for the specific activities throughout the system lifecycle." (Zimmerman, 2015c)

Both the development of technological assets and human competencies will be necessary to achieve these desired capabilities. While specific applications of model-centric engineering are continuing to develop and emerge, model-centric ideals and practices have begun to accelerate within certain fields (Zimmerman, 2015b).

The Office of the Deputy Assistant Secretary of Defense for Systems Engineering (ODASD(SE)) represents a key stakeholder in the development and practice of model-centric engineering. As the ideas and practice of model-centric engineering grow and evolve, the vocabulary used its discussion has also changed. For its purposes, the ODASD(SE) uses the following terms in its discussion of model-centric engineering:

> Digital Engineering: An integrated digital approach that uses authoritative sources of systems' data and models as a continuum across disciplines to support lifecycle activities from concept through disposal. (ODASD(SE), 2017)

> Digital Engineering Ecosystem**:** The interconnected infrastructure, environment, and methodology (process, methods, and tools) used to store, access, analyze, and visualize evolving systems' data and models to address the needs of the stakeholders. (ODASD(SE), 2017)

> Digital Model-Centric Engineering (DMCE): The application of engineering practices through the use of digital environments and tools. DMCE enables practitioners to engineer systems using digital practices and artifacts in a collaborative environment, creating a digitally integrated approach using a federated single source of truth to evolve complex systems. A primary characteristic of this environment and approach is the digital authority's ability to capture pedigree of all system-related data to facilitate and automate traceability, show dynamic relationships and changes to various aspects of the system development, and support decision makers to make informed decisions. (ODASD(SE), 2017)

This thesis' use of "model-centric engineering" is essentially synonymous with the ODASD(SE)'s use of "digital engineering" and "digital model-centric engineering." Similarly, "model-centric environments" are equivalent to "digital engineering ecosystems."

1.3 Scope

The term "model" means different things to different people. When I told my grandfather I was studying models in graduate school, he quipped that it sounds like quite an enjoyable subject to study (obviously referring to the fashion, rather than engineering, variety). For the purposes of this thesis, as stated earlier, models are an abstractions of reality that can come in a variety of forms and formats, but fundamentally are encapsulations of reality that humans use to augment their ability to make sense of the world, anticipate future outcomes, and make decisions. The U.S. Department of Defense Modeling and Simulation (M&S) Glossary defines a model as "[a] physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, or process." Examples of physical models include matchbox cars and a classroom globe – each physically represent a system, but are fundamentally simplified versions of the system. Models can be further abstracted to where they are no longer tangible physical models, but are instead descriptive and analytical models. Descriptive models describe, in various manners and with various levels of fidelity, logical relationships within a system ("Types of Models"). These descriptions may range from a complex system architecture view to a simple conceptual description on a PowerPoint slide. Analytical models are based on mathematical relationships that enable quantitative analysis through mathematical equations – physics-based models are a common form of analytical models ("Types of Models"). These are all forms of explicit models; however, one might also view human thought-processes as based on implicit mental models that structure our limited understandings of the world (M&S Glossary). In fact, modeling can be thought of as the process of making implicit mental models into explicit physical, descriptive, or analytical models. In order to scope this thesis appropriately around model definitions related to model-centric engineering, unless prefaced otherwise, the term model will broadly refer to forms of descriptive and analytical models that inform decision-making.

1.4 Research Questions

As stated in Blackburn et al. (2017), the "path forward" for model-centric engineering "has challenges but also many opportunities, both technical and sociotechnical." This thesis aims to generate greater insight into the sociotechnical aspect of model-centric engineering, specifically for human interaction with models within model-centric environments. The following research questions represent the core of what this thesis aims to accomplish:

Research Question 1: What human-model interaction challenges exist for individuals placed within model-centric environments, and how might they be mitigated?

Research Question 2: What technological and social factors exist that influence the trust and use of models in a decision-making process?

Research Question 3: What considerations are relevant for the incorporation of artificial intelligence (AI) and autonomy within model-centric engineering?

1.5 Key Contributions

This thesis presents the following key contributions to the study of human-model interaction and model-centric decision-making:

- **Analogy case studies.** Selected case studies that reveal relevant individual and organizational challenges to effective human-model interaction and decision-making within model-centric environments.

- **Expert interview investigation.** Empirical insight into sociotechnical factors that influence the trust and use of models by various types of actors within the model-centric decision-making process.
- **Considerations for AI and autonomy.** Relevant considerations for how the capabilities of AI and autonomy may relate to a model-centric context.
- **Heuristics**. Descriptive and prescriptive encapsulations of effective human-model interaction and model-centric decision-making to guide future practice, policy, and research. Following further validation, these have the potential to be fundamental principles for use in educating students and the existing workforce.
- **Policy considerations.** Recommendation to pursue a strategy of planned adaption that expands the policy-making process to one with established pathways for reducing uncertainty and iteratively creating more effective policies and guidance for model-centric engineering. Specific policy considerations based on this research are also provided.

1.6 Thesis Overview

This introduction serves to frame the discussion of this thesis within the model of the "4 C's" (see Figure 1). "Capabilities" represent the vision for what model-centric engineering hopes to achieve. "Competencies" are the knowledge, skills, abilities, and enabling technologies required to achieve the goals of the capabilities. To preempt naïve pursuit of these goals, however, empirically-based "cautions" serve to ground research in the reality of challenges that may hinder effective development if not appropriately addressed. Cautions then feed into "controls" created by policy-makers in order to iteratively feed into more nuanced and mature visions for the capabilities and competencies. The vision and motivation for the capabilities and competencies of model-centric engineering have been broadly articulated in this introduction; the rest of the thesis primarily focuses on research focused on cautions for the purpose of informing appropriate and effective controls.

Chapter 2 of this thesis details the methodology for how this research was conducted. Overall, the research can be described as exploratory research seeking to generate insight into a relatively unexplored area, rather than confirmatory, hypothesis-testing research. The Grounded Theory Method was the primary means for gathering and analyzing relevant qualitative data, primarily from expert interviews and existing literature.

Chapter 3 presents analogy case studies examining the cognitive and perceptual challenges faced by humans within complex decision-making environments where decisions must be made based on abstracted information. These case studies compile relevant empirically-based, descriptive evidence that can be used for identifying potential challenges that may be faced by humans within model-centric environments. The challenges are combined with further empirical and normative research aimed at developing potential mitigations of these issues. Additionally, a section on decision-making theory offers a model for understanding how and why biases occur, and how to mitigate their effects.

Chapter 4 presents the findings of expert interviews conducted for the sake of understanding how models and model-generated information are perceived, trusted, and ultimately used in a decision-making process. This study offers practitioner insight into human-model interaction, specifically concerning how models are used and trusted by various actors within the decision-making process. Findings indicate there are various factors and challenges that affect appropriate trust and use of models and model-generated information.

Chapter 5 presents a discussion on AI and autonomy, and how those capabilities may relate to a model-centric context. Model-centric environments will inevitably become more complex as they evolve to

capture the complexity of the systems they strive to model. Automation, autonomy, and artificial intelligence (AI) will likely play a key role in successful model-centric engineering. A fallacy of AI and autonomy is that such capabilities replace human interaction and control. This chapter argues that instead of replacing human will and intent, autonomous capabilities rather displace this human control, which necessarily changes how humans interact with those systems. Additionally, rather viewing intelligence as an individual attribute, decisions involving the balance of AI and human control should seek to extend the intelligence of the organization as a whole. As such, a discussion on AI and autonomy is timely and relevant to this thesis' overall goal of addressing human-model interaction within the context of model-centric engineering.

Chapter 6 presents an overarching discussion that combines research from previous chapters into guiding heuristics for the design, development, and use of model-centric practices. These heuristics represent proposed theories for human-model interaction, grounded in the empirical evidence and discussion offered through this research.

Chapter 7 proposes taking a strategy of planned adaptation for policy-making related to model-centric design, development, and use. The sociotechnical aspect of model-centric engineering coupled with uncertainty surrounding its development and use in the future creates a policy-space suitable for the strategy of planned regulatory adaptation. In addition to planned adaptation, more specific policy considerations are proposed based on the presented research.

Chapter 8 concludes with a summary of the research that revisits and answers the research questions. In addition, research limitations are discussed and recommendations are offered for potential future research to continue to advance the knowledge of human-model interaction.

2 Research Approach

This chapter provides an explanation of the approach taken in conducting this research. In addition to presenting a description of this research and its methodology, a section on research design lays out the various facets of this thesis, and how they contribute to the overall thesis goals.

2.1 Exploratory Research

This research seeks to generate insight into a relatively unexplored topic, specifically, the topic of human-model interaction. Taken in such a light, this research may be deemed *exploratory research*. According to Vogt and Johnson (2015), exploratory research is "research that looks for patterns, ideas, or hypotheses, rather than research that tries to test or confirm hypotheses." This definition takes an ambiguous term and makes it slightly less ambiguous by separating it from experiment-based research which starts with a hypothesis, and then seeks to confirm or disprove said hypothesis. As humans are of primary consideration in this thesis, this research could be further constrained to *social science exploration*, a term to which Stebbins (2001) offers the following definition:

> *Social science exploration* is a broad-ranging, purposive, systematic, prearranged undertaking designed to maximize the discovery of generalizations leading to descriptions and understanding of an area of social or psychological life (Stebbins, 2001).

Applying that definition, this research aims to use an intentional and systematic means for discovering generalizations concerning human-model interaction.

The goal of exploratory research is "development of theory from data" (Stebbins, 2001). Exploratory research involves a process of "inductively deriv[ing] generalizations about [a] group, process, activity, or situation under study" (Stebbins, 2001). These generalizations can be then formed into a *grounded theory* serving to explain something about the object of study (Stebbins, 2001). The term "grounded theory" has emerged as both an outcome of research – namely, theory grounded in data – as well as a method for developing that grounded theory. Inductively derived theory is not the end goal of research, but rather a critical piece within the larger process of research. Figure 2 illustrates how inductive, exploratory research of "little-known phenomena" provides the basis for increasingly predictive and confirmatory deductive research as the phenomena becomes more well-known and understood. Seen in this light, this research can be labeled as "qualitative-exploratory" research.

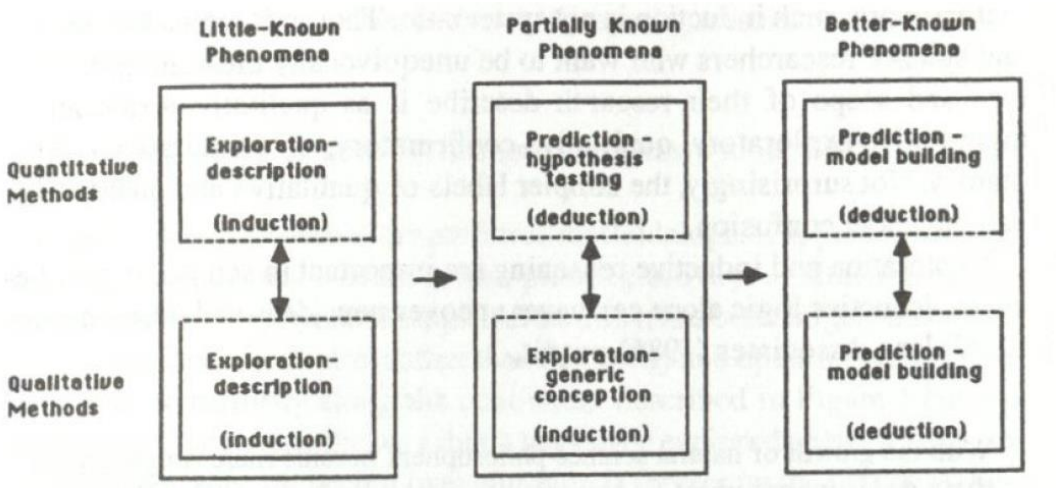


Figure 2. Relationship of qualitative and quantitative methods (source: Stebbins, 2001)

The hope for this exploration is that any discovered generalizations may prove a useful tool for better understanding the realm of human-model interaction, and in a manner that supports and promotes further exploration and research into this area. Exploration is a process, and this thesis does not intend to make claims of absolute completeness and finality, but rather, to present generalizations that can be built upon and expanded through continued research. Indeed, this is meant to be a process of *concatenation*, where each new data point of research forms a link in the chain of a clearer and more robust understanding of this research space. Within the constraints of time and resources surrounding any research project there will undoubtedly be areas of weakness in "sampling, validity, and generalizability," however, concatenation can smooth over and correct weaknesses as subsequent "links" of data points and research are added (Stebbins, 2001). Similar to how all models have inherent limitations and uncertainties, this thesis' encapsulation of human-model interaction undoubtedly has its own limitations that future research can improve upon.

2.1.1 Grounded Theory

The goal of exploratory research is grounded theory (GT), and the means towards attaining this goal is through a form of the grounded theory method (GTM). The foundational text of grounded theory, Glaser and Strauss' *The Discovery of Grounded Theory*, defines grounded theory as "the discovery of theory from data systematically obtained from social research" (Glaser and Strauss, 1967). This theory generation "involves a process of research" where "most hypotheses and concepts not only come from the data, but are systematically worked out in relation to the data during the course of the research" (Glaser and Strauss, 1967). This stands in contrast to "theories logically deduced from *a priori* assumptions" (Glaser and Strauss, 1967). In contrast to research relegated to testing and verifying existing theories, *The Discovery of Grounded Theory* provides a seminal foundation for advancing the importance of developing one's own theories from data. In order to obtain grounded theory, *The Discovery of Grounded Theory* proposes various methods for developing grounded theory. Therefore, "grounded theory" is the overall goal and result of the grounded theory method (Bryant and Charmaz, 2007).

While *The Discovery of Grounded Theory* provides a general grounded theory method, various methods for producing grounded theory have evolved since its 1967 publication. Numerous additional authors have contributed to GTM development, and "the method itself has now taken on a life of its own" (Bryant and Charmaz, 2007). Various perspectives and points of contention have emerged, including between the original founders themselves. While many elements of grounded theory method may be similar, variations have emerged in the execution of this method, from specific, prescriptive, "cookbook" type approaches, to more broad and general methods of theory development. Rather than focusing on prescriptive rules, however, GTM is intended to be "based around heuristics and guidelines (Bryant and Charmaz, 2007). Rather than concerning ourselves with the history of GTM development and many approaches towards obtaining GT, of which numerous publications can be found (Bryant and Charmaz, 2007; Covan, 2007; Star, 2007), this methods section will concern itself with the specific aspects of the GTM employed in this research.

Theoretical Sampling

Before theory can be developed from, and grounded in, data, the data must be collected. As data provides the foundation for grounded theory, the process of collecting data proves critical to the outcome of the research. Glasser and Strauss (1967) labels this process of data collection as theoretical sampling (i.e. sampling to collect data for generating theory, not hypothetical, "theoretically speaking," sampling that does not actually take place). Theoretical sampling is defined as "the process of data collection for generating theory whereby the analyst jointly collects, codes, and analyzes his data and decides what data to collect next and where to find them, in order to develop his theory as it emerges" (Glaser and Strauss,

1967). Rather than preconceived notions and agendas dictating how data is collected, the process of data collection should be "controlled by the emerging theory" (Glaser and Strauss, 1967).

A general critique about the GTM concerns its subjective nature and the potential for researcher bias in both data collection and data analysis. This is a valid critique, one that applies more strongly to qualitative research than quantitative research, as qualitative research requires a greater amount of interpretation. This need for interpretation is unavoidable, however, and as Strauss and Corbin (1998) notes, "at least it is interpretation based on systematically carried out inquiry."

Comparative Analysis and Coding

In short, developing theory through the GTM involves "developing abstract concepts and specifying the relations between them" (Strauss and Corbin, 1998). Many grounded theory methods use the term "coding" as the process for grouping collected data and developing these concepts. According to Strauss and Corbin (1998), coding is "the analytic [process] through which data are fractured, conceptualized, and integrated to form theory." In the modern age of computers, the term "coding" might cause some confusion as it can imply the process of writing code for software, however, as used in the GTM, and for the purposes of this thesis, "coding" is referred to as the process of systematically categorizing data into categories and concepts from which relations and theory can later be derived. Therefore, this thesis employs a process of "captur[ing] patterns and themes and cluster[ing] them under a 'title'" for the purpose of aiding analysis and interpretation of data (Strauss and Corbin, 1998). Strauss and Corbin (1998) emphasizes the difference between description and theory in GT research. GT research seeks to move past mere description of an area of study and on to theory; "theory" defined by Strauss and Corbin (1998) is "[a] set of well-developed concepts related through statements of relationship, which together constitute an integrated framework that can be used to explain or predict phenomena." Therefore, coding provides a tool for theorizing by systematically creating categories and concepts from the data. This data placed within the developed categories and concepts can then be analyzed for identifying relationships that offer an explanation about phenomena; an explanation that moves past mere description (Strauss and Corbin, 1998).

Sources of Data

The data generated and used in this research stems from two primary sources: expert interviews and existing literature. Literature is used throughout this thesis to not only provide background information, but also to provide additional data points for the formation of theory. The existing literature referenced throughout, provides a wealth of data used to inform and develop the theories presented. In the pre-Internet timeframe when *The Discovery of Grounded Theory* was published, Glasser and Strauss emphasize the importance and relevance of "library research." They argue that "when someone stands in the library stack, he is, metaphorically, surrounded by voices begging to be heard" (Glaser and Strauss, 1967). Through the use of the Internet in this current age, these voices are abundantly more accessible, and "[t]he researcher needs only to discover the voices in the library [and Internet] to release them for his analytic use" (Glaser and Strauss, 1967).

The other main source of data for this research comes directly from people. The majority of this data was collected through semi-structured interviews with thirty domain experts. The data from these interviews and conversations were captured primarily through note taking, recordings and transcriptions (from those who consented to recording). The research protocol was approved by MIT's Committee on the Use of Humans as Experimental Subjects (COUHES) and a DoD Internal Review Board (IRB).

Theoretical Saturation

Key to grounded theory research is that the theory emerges from data, however, one challenge with exploratory grounded theory research is knowing when to stop collecting data. As with any research project, scoping is necessary for ensuring the project can be completed within a reasonable timeframe. Although exploratory in nature, this research is purposefully bounded by its key objectives and research questions. Research is a process, which allows the questions and data collection to change and evolve through the emergence of insights from data already collected. However, if the project is appropriately scoped, eventually fewer novel insights will be gained through additional sampling and data collection. When developing theory, the general rule is to gather data until "each category is saturated" (Strauss and Corbin, 1998). This theoretical saturation is obtained when categories are well developed and "no new or relevant data seem to emerge" within the categories (Strauss and Corbin, 1998). Such a determination relies upon researcher judgment and carries inherent limitations. Undoubtedly, carrying on the study indefinitely is the only way to be sure that all relevant data is collected, however, for the sake of feasibility, the strategy of judging a sufficient saturation point aligns appropriately with the goals of this research.

2.2 Research Design

The previous section provides an explanation of GT, the GTM, and how this thesis employs those concepts and ideas in pursuit of generating grounded theory. The following section addresses the design of this research, and how various elements combine to achieve the overarching research objectives.

2.2.1 Descriptive, Normative, Prescriptive Goals

The MIT Systems Engineering Advancement Research Initiative (SEAri) research model (Figure 3) emphasizes three key areas of research in the pursuit of effective solutions: descriptive, normative, and prescriptive research. In the process of better understanding the problem space of human-model interaction, this thesis aims to provide heuristics and policy guidance to aid the development and use of model-centric engineering practices. These heuristics are both prescriptive and descriptive theories generated through this research to guide model-centric practice and policy, and therefore, represent outputs related to the goal of prescriptive research within this thesis. This prescriptive research, however, builds off of descriptive and normative research. The descriptive research in this thesis examines how human actually make decisions and interact with models and other relevant complex technologies. The expert interview study provides primary-source information through interviewing and conversing with expert practitioners. Case studies examine published research to gather information describing humans-interaction with complex systems like glass cockpits and nuclear reactors While descriptive research examines how people *actually* perform, normative research seeks to understand how individual *ought* to behave. Literature on decision-making, biases, and associated mitigations provide theory-based, normative research to inform and more fully develop what is learned through the descriptive interviews and case studies. The descriptive and normative threads ultimately blend together to produce the prescriptive outputs of heuristics for how to handle various aspects of human-model interaction and model-centric decision-making. As indicated by Figure 3, the process of prescriptive, descriptive, and normative research should cyclically continue in an effort to gain greater insight into an area of study, of which this thesis only scrapes the surface.

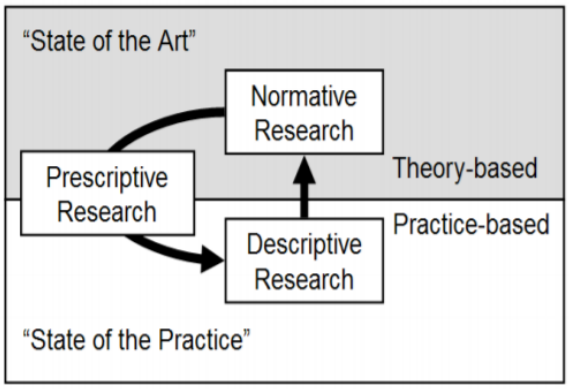


Figure 3. MIT SEAri Research Model (source: Ross and Rhodes, 2008)

2.2.2 Literature-Based Data

Rather than solely providing a backdrop and introduction to research, literature may also constitute a source of data and knowledge integrated towards the advancement of the research. Strauss and Corbin (1998) provide various ways in which existing literature may be used as valid data within the Grounded Theory Method:

- Literature can provide concepts used for making comparisons with other collected data while developing theory through comparative analysis.
- Although one does not want literature to stifle the emergence of theory through preconceived notions, a researcher familiar with relevant literature may be more sensitive to "subtle nuances in data."
- Descriptions, quotations and perspectives found within published literature may be used as a secondary source of data.
- Literature may direct the research to consider relevant situations or questions appropriate to the research goals.
- Literature can be used for "extending, validating, and refining knowledge in the field."

Existing literature not only provides a means for justifying the relevancy of research, but can also offer an integral portion of data used in the analysis and development of the research as a whole, as is so with this thesis. A separate literature review chapter is not included within this thesis, rather, relevant literature is provided throughout the its various chapters.

2.2.3 Analogy Case Studies

While this thesis aims to generate insight to inform model-centric development and use, little empirical data exists concerning human-interaction in model-centric environments. While not strictly modeling environments, however, humans do operate within various other complex technological environments in which they have to make decisions based off of abstracted information. This thesis examines empirical research of human-interaction in two different environments: aircraft glass cockpits, and nuclear reactors. While these environments differ, humans provide the common factor that ties them together, and lessons learned in these environments provide insight for human considerations of model-centric environments.

2.2.4 Decision-Making Theory

A literature review on dual-process theory offers a model for understanding individual decision-making and where it can go wrong. As effective decision-making is a primary goal of model-centric engineering,

biases that lead to erroneous judgment are undesirable. This theory helps move past simply describing biases and erroneous judgment, and on to presenting an explanation for why they happen. With such an understanding, steps can then be taken to mitigate the systematic and predictable biases that plague individual judgment and decision-making.

2.2.5 Expert Interviews

An empirically-based understanding of how individual actors and decision-makers within the current modeling landscape interact with, trust, and use models may help aid the transition to model-centric practices. The expert interviews conducted in this research seek to understand the process in which models are used to influence decisions, along with what factors affect this process. There are many different types of actors involved in this process, so this specific study aims to understand the technological and social factors that affect those various actors and final decisions.

3 Analogy Case Studies and Decision-Making Theory

Relatively little information on human-model interaction in model-centric environments exists, and this chapter presents discussions on the following topics in order to advance knowledge in that area: pilot interaction with glass cockpits, nuclear reactor operators, and dual-process theory of cognition and decision-making These discussions draw on existing descriptive research (how people act) and normative research (how people ought to act) in specific contexts for the purpose of drawing out general principles that can be used to inform human-model interaction and decision-making with model-centric engineering. The glass cockpit study introduces various cognitive and perceptual challenges that can plague individual decision-makers in model-centric-type environments. In addition to expanding on individual challenges, the nuclear reactor case study broadens the scope of analysis by identifying various organizational factors that can ultimately influence improper individual behavior. Finally, the section on decision-making presents a model of dual-process theory to help explain where biased judgments come from and how they might be mitigated.

3.1 Analogy Case Study: Glass Cockpit

This case study focuses on the cognitive and perceptual aspects of intensive human-model interaction, and explores relevant findings and lessons learned from the experience of aircraft pilots with glass cockpits and virtual displays. This thesis postulates that relevant similarities exist for system designers and decision-makers within immersive model-centric environments, with increased automation, interactivity and abstraction of systems information.

3.1.1 Introduction

Rhodes and Ross (2015) expresses that models "represent an abstraction of reality," and "can come in a variety of forms and formats, but fundamentally they are an encapsulation of reality that humans use to augment their ability to make sense of the world and anticipate future outcomes." The idea that "humans use" models highlights human interaction as a necessary factor for all models. Given this common characteristic of human interaction, this thesis proposes that experiences gained in one model-centric situation can offer insight into entirely different model-centric environments.

Aircraft cockpit displays present pilots with models of the aircraft's state in order to facilitate appropriate decision making and action. This case study explores the experience of aircraft pilots with digital "glass cockpit" displays in an effort to draw out lessons learned from the glass cockpit's impact on human and system performance. Through analysis of aircraft accidents and subsequent research findings, areas of concern in the interaction between glass cockpits and human pilots are identified. Lessons are drawn from substantial research that has been conducted to not only retroactively address accidents, but also to identify areas susceptible to failure and to determine the causes of these failures, with an end goal of mitigating future occurrences of accidents. Operating on the premise that the cognitive and perceptual issues found in the cockpit transcend to broader terms of human-model interaction, this investigation explores these lessons in order to spark discussion and thought into the role of human-model interaction within the emerging field of interactive model-centric engineering.

3.1.2 Glass Cockpits and Automation Related Accidents

The term "glass cockpit" began making its way into the aviation community in the 1970s with the transition from the use of electromechanical instruments to electronic flight displays. Used initially to describe displays incorporating cathode ray tubes, "glass cockpit" has since evolved into a descriptor for digital flight displays and automation systems within aircraft in general (see Figure 4) (*Introduction of Glass Cockpit Avionics).* The arrival of the glass cockpit equipped Boeing 757 and 767 in the early 1980s ushered

in the use of glass cockpit and automation technology within commercial aviation, progressing to become standard design in nearly all modern aircraft (Strauch, 1998). This new technology sought to improve system functionality by increasing human capability and efficiency through automation of flight operations and ultimately allowed the crew composition of commercial aircraft to be reduced from three to two members (Wiener, 1989). As noted by Endsley (1996), however, these benefits from automation also accompanied changing the pilot's role from flying to monitoring an automated system, a "role people are not ideally suited to." Analysis of aircraft accident case studies provides insight into challenges that glass cockpits and associated automation have caused for pilots within the cockpit environment.

Figure 4. (a) traditional dial cockpit; (b) glass cockpit.

Nagoya Accident

On April 26th, 1994, while piloting an Airbus A300-600 on landing approach in Nagoya, Japan, the First Officer (FO) mistakenly engaged the Go-Around mode as the aircraft neared 1000ft above ground level. The aircraft appropriately responded by autonomously adding power and initiating a climb that the FO tried to manually counteract in order to keep the plane on the appropriate glide path. While the Captain noticed the erroneous initiation of the Go-Around mode and told the FO to disengage the mode, the FO failed to do so. The FO managed to halt the plane's ascent and engaged the autopilot; 19 seconds later, however, the autopilot caused the plane to pitch up again and the FO subsequently disengaged the autopilot. Around 570ft, the aircraft sensed near-stall conditions and autonomously staged a stall recovery which began a climb once again. This time the pilots were unable to stop the climb which ultimately led to a stall, inadequate time for recovery, and a tail-first crash landing killing 264 of the 271 individuals onboard (Baxter, Besnard, and Riley, 2010).

Strasbourg Accident

The next case example occurred in Strasbourg, France on January 20, 1992 when an Airbus 320 impacted the ground while on descent for landing. Prior to beginning the approach, the crew received last minute instructions from Air Traffic Control to complete a straight in landing rather than the expected circling approach. This unanticipated guidance resulted in an increased workload for the crew as they worked to complete the preparations for landing in an earlier than expected manner. As part of the preparations, the crew entered the number "33" into the flight computer to set the appropriate glide path angle of -3.3 degrees. They failed to realize, however, that the computer's mode was set to rate of descent and that they actually commanded the aircraft to descend at 3,300 ft/min. The aircraft proceeded as was mistakenly directed and subsequently crashed into the ground well short of the runway, with only nine of the 96 individuals onboard surviving (Strauch, 1998; Mosier, Sethi, McCauley, Khoo, and Orasnu, 2007).

Cali Accident

Reminiscent to the event at Strasbourg, an American Airlines Boeing B757 received Air Traffic Control guidance on December 20, 1995 to complete an unplanned, straight-in landing approach for its destination, Cali, Colombia. Needing to adjust their flight plan to complete the approach, the crew proceeded to enter in the next appropriate navigation waypoint, "ROZO," into the flight computer. After inputting "RO," however, the waypoint "ROMEO" was the first available point on the list which the crew mistakenly selected; the aircraft then began navigating to a waypoint located 132 miles away from the destination. Approximately a minute following the plane's course adjustment away from Cali, the crew realized their mistake and reprogrammed the flight to the appropriate point, ROZO. Assuming the situation rectified, the crew failed to realize the deviation from the original flight path set the airliner on a collision course with a mountainside. Only 4 individuals out of 163 survived the crash (Strauch, 1998; Besnard and Baxter, 2006).

Human-Automation Breakdown

The three accident examples underline a similar theme in that they likely would not have occurred in the absence of highly automated equipment and integrated displays within the cockpit – all demonstrating a breakdown of human interactivity with the aircraft that ultimately led to devastating results. This evidence opens the door to questions concerning the causes and potential mitigations of these errors along with presenting an opportunity to gain insight into human-model interactivity, specifically in highly automated environments like those found in aircraft.

3.1.3 Cognitive Coherence

The transition of aircraft cockpit technology has largely changed the role of the pilot from one that requires "stick-and-rudder" skills, to one primarily concerned with programming and monitoring the aircraft's automation (Mosier, Skitka, and McDonnell, 2001). As described by Mosier et al. (2001), this shift in the pilot's role also accentuates the importance of coherence competence: "an individual's ability to maintain logical consistency in diagnoses, judgments, or decisions." The displays within aircraft present nearly all of the necessary data to safely fly the plane, and if the pilot can maintain coherence and take appropriate action throughout the entirety of the flight then the pilot has succeeded. Mosier et al. (2007) also notes that many piloting errors manifest themselves as failures of coherence in that they fail "to note or analyze important information in the electronic 'story' that is not consistent with the rest of the picture." The outcomes of the previous cases all resulted from a failure to maintain coherence throughout the entire flight. While maintaining coherence is a primary objective for pilots, there are many means through which automation can contribute to the breakdown of effective coherence.

3.1.4 Automation Bias

Mosier and Skitka (1999) defines automation bias as "the use of automation as a heuristic replacement for vigilant information seeking and processing," which can result in commission errors (incorrectly following an unverified automation directive) and omission errors (failing to identify an issue not identified by an autonomous system). An everyday example of a commission error would be a driver blindly following a GPS navigation aid's incorrect directive to turn the wrong way onto a one-way street. Additionally, missing the proper highway exit due to lack of warning from the navigation system would constitute an error of omission (Parasuraman and Manzey, 2010). Specifically related to human interaction with automated decision aids, automation bias seems to be influenced by three different factors. First, humans often choose to proceed down the path of least cognitive effort. This can lead to using automated aids as strong decision making heuristics while failing to seek out all relevant information to develop the full, coherent picture of the situation. Humans also exhibit a tendency to perceive automated decision making and performance as superior to their own, leading to an overestimated trust that the system is performing appropriately for the

given situation. A third factor influencing automation bias is the phenomena of perceiving automated aids as fellow crew members and diffusing responsibility. This can lead to a "social loafing" behavior where human operators perceive themselves as less responsible for the system performance and outcome (Parasuraman and Manzey, 2010).

The accidents at Strasbourg and Cali offer examples that manifest potential instances of automation bias. At Strasbourg, after the initial mistake of entering the data as a descent rate rather than flight path angle, the crew failed to vigilantly validate the aircraft's descent against other relevant forms of information, thus committing an error of omission. In the landing approach to Cali, the flight computer suggested the incorrect waypoint, ROMEO, and the crew committed a commission error by blindly following the automated suggestion and not adequately processing the information they received.

3.1.5 Automation-Induced Complacency

Definitions of complacency include: "self-satisfaction that may result in non-vigilance based on an unjustified assumption of satisfactory system state," and "a psychological state characterized by a low index of suspicion" (Parasuraman and Manzey, 2010). Pertaining to aviation, one can readily imagine the negative impacts pilot complacency can have on the safety of flight; in fact, an early 1970's study by NASA on the effects of automation in the cockpit identified complacency as a key area of concern for pilots when questioned on their perspective on automation's potential impact on safety (Wiener, 1989). Research by Parasuraman and Manzey (2010) goes on to define automation-induced complacency as "poorer detection of system malfunctions under automation control compared with manual control." This failure in achieving a fully coherent picture typically manifests itself under periods of high, multi-task work load, and constitutes an active diversion of attention from automation to other manual tasks (Parasuraman and Manzey, 2010). While this relocation of attention resources may be an understandable reaction of pilots under high workloads, it is by no means an acceptable response as it is the pilot's job to remain aware of all relevant information and processes, and failure to do so can produce devastating results. Although readily understood and accepted as undesirable, automation-induced complacency presents a challenge in that complacent behavior may seldom produce negative results since systems typically operate as expected. This can lead to failure of awareness and even possible acceptance of the behavior. In highly intensive and unforgiving systems like aircraft, however, all it can take is one unnoticed failure for there to be grave consequences.

Automation-induced complacency is closely related to automation bias as they both present manifestations of similar attentional issues. Most similarly, both automation-induced complacency and automation bias can result in errors of omission. Automation-induced complacency can result in this error from failure to appropriately monitor the automation itself due to diversion of attention, while automation bias results in failure to adequately monitor the system as a whole due to a bias that the automation will warn the operator if something goes wrong. All the case examples appear to exhibit complacent behavior to some degree. In Nagoya, the FO's mistake of engaging the Go-Around mode could have been an innocent mistake, but both his failure to appropriately monitor the automation and fix the error along with the captain's failure to ensure situation rectified lend themselves to complacent behavior. Both Strasbourg and Cali also show examples of incorrectly assumed satisfactory state of the system and automation although non-complacent behavior likely would have detected the mistakes in time.

3.1.6 Mode Error

Modes serve as a means through which automation can extend human capability by structuring complexity and presenting users with varying levels of control styles (i.e. "modes" of operation) (Chappell, Crowther,

Mitchell, and Govindaraj, 1997). Glass cockpits have capitalized on the use of modes by giving pilots means to tailor the aircraft's automation to specific situations and preferences (Chappell et al., 1997). Yet, as with most technology, new capabilities are closely paired with new pathways to potential failure. Specific to modes, a breakdown in coherence can occur when the human operator "loses track of which mode the device is in" (Sarter and Woods, 1992). Known as mode error, this breakdown results in a misinterpretation of the situation and unwanted system responses to given inputs (Sarter and Woods, 1992). Research suggests that mode error occurs through a combination of "gaps and misconceptions" in operators' model of the automated systems and the failure of the automation interface to provide users with salient indications of its status and behavior (Sarter, Woods, and Billings, 1997). This propensity for lack of mode awareness in glass cockpits was accentuated by a NASA study in 1989 where 55% of pilots encountered automation surprises after more than one year of flying in glass cockpit aircraft (Wiener, 1989; Sarter et al., 1997).

Indeed, the Strasbourg accident clearly shows a crew committing mode error by failing to realize that they entered "33" into the descent rate mode rather than the desired descent angle mode. Had the crew maintained the proper awareness of the system's actual mode, they would have switched to the proper flight path angle mode without an issue and avoided their deadly error. Similarly, at Nagoya, the aircraft responded appropriately given the Go-Around mode that was inadvertently commanded, yet the crew failed to understand the response of the aircraft and how to appropriately handle it, which ultimately led to the crash.

3.1.7 Perceptual Challenges

A successful system design must not solely take into account how information is cognitively processed, but also how information is perceived. The previous case examples have shown areas where glass cockpit technologies can contribute to cognitive failures, but additional research on the transition from analog to glass has revealed areas of perceptual failure. This aspect of perception must also be addressed in the design of an effective system.

Human-Machine Interface

From a performance point of view, a simple question can be asked when discussing analog and glass cockpits: which results in better performance? The purpose behind transitioning to glass was not only to make the pilot's job easier through increased automation, but also to enhance the performance and safety of the aircraft overall. While glass cockpits have undoubtedly provided benefits in many aspects of flight, they do not necessarily yield better performance in all areas. A study by Wright and O'Hare (2015) compares simulator flight performance between participants using traditional analog instruments and those using advanced glass cockpit displays, specifically comparing performance in loss of control events, and accuracy in maintaining altitude, airspeed, and heading. The results show that the traditional cockpits actually resulted in better overall performance, corroborating with a separate study conducted by Hiremath, Proctor, Fanjoy, Feyen, and Young (2009) which demonstrates that glass cockpit users had longer recovery times from unusual attitude situations than traditional cockpit users. One explanation for this disparity stems from the manner in which relevant information (airspeed, altitude, attitude, etc.) is presented and received. Traditional cockpits use individual round dials with indicator needles for each piece of flight information, while glass cockpits integrate much of the data into a computer display and present airspeed and altitude as a moving tape with an exact readout (see Figure 4). Dial instruments offer a means for obtaining information at a glance by allowing pilots to see where the needle is in relation to the whole range of numbers rather than requiring an exact readout as found on glass displays (Hiremath et al., 2009). This ability to take in information at a glance allows the pilot to more quickly assess the state of the aircraft and adjust accordingly. Safe piloting does not necessarily require adherence to an exact number, as it is more important

to stay within an appropriate range of numbers. Glass cockpits do not include the entire range of numbers which makes it harder to discern if the aircraft is in the appropriate range. These studies indicate that system designers must not only understand what information must be presented to users, but also understand how to present the information in a manner that most effectively accomplishes the tasks at hand.

Preference-Performance Dissociation

An important factor to consider in evaluating performance is not merely how the user objectively performs, but also how the user *thinks* he or she is performing. In the Wright and O'Hare (2015) study, the pilot test participants unanimously rated the glass cockpit superior to the traditional display. In their perception, the glass cockpit offered the "most awareness-enhancing, the least mentally demanding, and the easiest to interpret" display with the "fewest disliked features." Despite this perceived superiority, the pilots actually performed worse with the display they preferred the most. This highlights a phenomenon known as "preference performance dissociation" where users' preferences do not line up with their performance (Andre and Wickens, 1995). In the case of glass cockpits, Wright and O'Hare (2015) postulates that simply the use of bright, highly contrasted colors results in the superior feedback as humans have been shown to prefer color as opposed to lack thereof. This presents a need to understand that users do not necessarily know what is best for them, and that sometimes "user-centered" design should involve designing for how the user actually performs, and not just for what the user wants.

3.1.8 Implications for Model-Centric Engineering

One of the most obvious differences between flying a plane and preforming engineering tasks is the criticality of time in relation to successful (or unsuccessful) results. In an aviation environment, many decisions and subsequent actions must be performed rapidly to ensure proper control and safety of the aircraft, with an increased time pressure workload demand shown to greatly affect performance and propensity for error (Mosier et al., 2007). Modeling environments, on the other hand, have much greater margins of freedom in regards to time required between recognition of a need and required action. Basically, modelers do not have the threat of imminent death if they take some time solving a problem. While decreasing time pressure can indeed help mitigate challenges of automation-related errors, it does not completely ameliorate the possibility of humans making mistakes while interacting with automation. Additionally, one can reasonably foresee modeling environments changing with technology for much of the same reasons as cockpits have to include: increased human capability, greater efficiency of operations and human resources, etc. The increased capabilities that automated and intensive modeling environments provide would not only allow for work to be accomplished at a quicker pace, but could also increase time-pressures for modelers, exacerbating the existing propensity for error.

Model-Centric Engineering and Automation Scenario

Imagine a new modeling environment like the one illustrated in Figure 4. This environment seeks to create greater efficiency and optimality in decision making by incorporating the benefits of increased modeling autonomy and multi-stakeholder collaboration. Decision makers in this environment can create and share real-time exploration and changes to models, facilitating the understanding of how design choices impact desired stakeholder preferences and system performance. In this example, your team is working on developing an overdue recommendation to determine what design the project will move forward with, and must arrive at a final solution by the end of the day. You and your fellow team members are separately working with different models to perform analysis tasks previously accomplished by multiple people over longer time. Now, automated decision aids and state of the art software help facilitate cross-model analysis and convergence, resulting in an interim outcome that reflects what you expect to see. Assuming the automated system accurately determined what you had requested, you accept the outcome at face value and

allow the model to continue further analysis toward a solution as you leave for a quick coffee break. While you are away you fail to realize that the model was operating, and continues to operate, in an incorrect mode. Never encountering any issues before and trusting in the system's high fidelity, you accept the model's flawed recommendation and share the result with your teammates who also consent to the recommendation without critique; there have been no problems before, plus the day is late and everyone wants to go home. While much quicker than previously possible, the decision making process was interwoven with automation bias, mode error, and complacent behavior under increased time pressure which results with settling upon a suboptimal design.

3.1.9 Mitigating the Challenges of Human-Automation Interaction

Up to this point, the case study has explored potential areas of failure and challenge in the realm of human-automation interaction with the end goal of extrapolating these lessons to human-model interaction. We have seen how the transition to glass cockpits fundamentally changed the role of the pilot in the cockpit, invariably creating new challenges that the pilot must overcome. This next section serves as an introductory exploration into potential means for mitigating the negative effects of human-automation interaction in the cockpit and implications for the model-centric workplace. By no means are these fully developed solutions, but rather they should serve as an introduction to potential guiding principles for developing interactive model-centric environments.

Accountability

The use of social accountability has been demonstrated as effective in mitigating various cognitive biases to include primacy effects, the fundamental attribution error, over-confidence effects, and the "sunk cost" effect (Skitka, Mosier, and Burdick, 2000). Taking this a step further, Skitka et al. (2000) goes on to test the efficacy of accountability in ameliorating the effects of automation bias. A study by Mosier, Skitka, Heers, and Burdick, (2009) found that pilots who "reported a higher internalized sense of accountability" verified correct automation functioning more often and committed fewer errors in the automated environment. This matches well with earlier research that found that properly channeled accountability results in mitigating automation bias through lower rates of automation bias related errors (Skitka et al., 2000). Automation bias presents a clear pathway of coherence breakdown in the human-model interaction, and the use of adding accountability offers a means for ameliorating this issue. While the means of assigning accountability can be varied in application and effectiveness, it is important to note that assignment and internalization of accountability can have important effects towards limiting automation bias. Little research is available that specifically addresses accountability and automation complacency, but the similarity between complacency and automation bias suggests a potential link between accountability and complacency as well.

Transparent Systems

Achieving coherent mode awareness, the ability to effectively understand, follow, and anticipate the behavior of automated systems, is the understandable solution for preventing mode error (Sarter et al., 1997). Mode error does not necessarily result from complacent or biased human behavior, but rather from the fact that users can sometimes fundamentally lose track of what mode the system is in. Specifically addressing aviation, Besnard and Baxter (2006) argues that the development of transparent flightdecks begins to address the issue of achieving mode awareness. Along this line of thought, modeling environments should enable transparency as needed and allow the user to understand, follow, and predict the automation's behavior. Designing for model transparency could contribute greatly to reducing automation surprises and subsequent errors by model users by offering increased insight into the actual functioning of the models.

Human-Centered Design

As long as human operators bear ultimate responsibility for system performance, they must be integrated into the system and provided with all relevant information needed to assess the system's performance, state, and behavior (Sarter et al., 1997). With this in mind, the challenge in design becomes one that focuses on the effective integration of the human into the system rather than forcing the human to adapt to the system; in other words, human-centered design is needed rather than technology-centered (Sarter et al., 1997). Some of the issues with modern automation in cockpits result from including technology simply because it is technically feasible (Besnard and Baxter, 2006). This approach can lead to a highly capable system, but very ineffective system if the human is not appropriately designed for. Norman emphasizes this paradigm shift in design thinking by rephrasing the motto of the 1933 Chicago World Fair from "Science finds, industry applies, man conforms" to the contrasting idea of "people propose, science studies, technology conforms" (Sarter et al., 1997). Continued focus on human-centered design within model-centric environments is needed in order to advance the compatibility, success, and effectiveness of human-automation model interactions.

3.1.10 Conclusion

As technology has evolved and developed in its capabilities, models have similarly progressed to include greater fidelity and functionality in the pursuit of more adeptly abstracting reality for the use of designers and decision-makers. Model-centric engineering stands as a developing practice that promises effective capabilities for efficiently realizing successful systems through intense human-model interaction. With innovative ideas and technologies, however, also come new sources of potential failure. As a means for sparking thought and discussion, this case study has presented the introduction of glass cockpits into aircraft as an analogy case for addressing potential issues to be faced in the transition from traditional engineering to the use of interactive model-centric environments.

The use of advanced technology in cockpits manifests itself primarily through an increase in autonomy that not only changes the role of pilots, but also adds an additional component: manager of systems (Besnard and Baxter, 2006). While this technology has been successfully integrated into modern aviation, it also highlights the continued importance of considering the human interaction with technology, as specifically evidenced by disastrous examples. The discussions on the cognitive coherence failures of automation bias, complacency, and mode error combined with perceptual areas of concern provide a starting point for educating model developers, model users and decision makers on ways that effective human interaction with model-centric technology is prone to failure. Offered not as fully developed solutions, but rather initial guiding principles for mitigation, the heuristics on the importance of accountability, transparency of systems, and human-centered design begin to address means for mitigating potential failure points and achieving greater effectiveness in the realm of model-centric engineering.

3.2 Analogy Case Study: Nuclear Reactor Operators

While aircraft cockpits represent one existing form of immersive environments where humans make decisions based on abstracted information, nuclear power plant (NPP) control rooms provide another insightful example. This case study examines the role of reactor operators within nuclear power plants. Two specific examples, Chernobyl and Three Mile Island, offer insight into various pathways contributing to decision-making and judgmental failure. These examples, coupled with additional literature on human factor challenges within NPP control rooms, supply further insight into individual and organizational challenges that could hinder effective model-centric decision-making.

3.2.1 Nuclear Reactor Operators

A nuclear reactor operator holds the responsibility of monitoring, controlling, and ensuring safe operation of the nuclear reactor at a NPP. At the heart of the operator's control is the ability to increase or decrease the reactivity of the reactor by inserting or removing neutron-absorbing control rods ("Control Rods"). Inserting control rods leads to greater absorption of neutrons, which in turn reduces the amount of fission reactions and subsequent power generation. Therefore, by operating the control rods, reactor operators can maintain the desired state of fission reactions. While the nuclear reactor itself represents the central system of a nuclear plant, numerous other systems, such as the cooling and power generating systems, work together in a complex interrelationship. As these various tangential, yet critically linked, plant systems are all necessary to achieve safe and effective reactor operation, reactor operators have the responsibility of monitoring and controlling them as well. The various tasks of a nuclear operator can be classified into four tasks (Bovell, Carter, and Beck, 1997):

- Function monitoring
- Fault diagnostic
- Control manipulation
- Administrative

Function monitoring includes monitoring the thousands of instruments and displays providing information about the various components that comprise the overall NPP system (Mumaw, Roth, Vicente, and Burns, 2000). Much like displayed information within glass cockpits, this information represents abstracted pieces of information from which operators must assess the states of the power plant. Auditory and visual alarms, also existing on the order of thousands, are used within the control room to assist operators in identifying areas needing attention and potential action.

Fault diagnostic tasks include the two-part process of analyzing problems and choosing best alternatives for corrective action (Bovell et al., 1997). Considering the thousands of displays and alarms found within a control room, the joint tasks of monitoring and fault diagnostics present large cognitive workloads. Monitoring and fault diagnostics also involve not only dealing with alarms and deviances when they arise, but also assessing and prioritizing the alarms that are nearly always present (Mumaw et al., 2000).

Operators may manipulate the control rods in response to preplanned changes to the power output, or to provide reactive corrections to the system, such as conducting an emergency shutdown to prevent a meltdown. Administrative activities provide additional workload activities such as log-keeping, event reporting, and system testing (Bovell et al., 1997). Given the risks involved with operating a NPP, operators must be highly trained and skilled. Errors do occur, however, with varying levels of consequence.

Figure 5. Example of Nuclear Power Plant Control Room (Copyright © 2009 Yovko Lambrev, Creative Commons)

3.2.2 Chernobyl

The Chernobyl NPP accident offers a vivid reminder of the potential consequences of poor system design, and improper human performance. Occurring in 1986 within the then Soviet-controlled Ukraine, the Chernobyl disaster is the only commercial nuclear power plant accident to cause direct fatalities from radiation ("Chernobyl Accident and Its Consequences"). In addition to the two personnel killed on-site by the explosion, the accident resulted in the deaths of twenty-eight individuals due to radiation exposure and thermal burns, and may have contributed to thousands of cases of thyroid cancer among the surrounding population in the years following ("Chernobyl Accident and Its Consequences").

The Accident

The accident occurred when operators inappropriately ran a scheduled test of the system. The Chernobyl reactor was not designed for sustained operation below 700 MW(th); however, the test was initiated by the operators at 200 MW(th) (INSAG-7). Due to its underpowered state, the reactor was highly unstable when the test began, which led to an accelerated nuclear chain reaction and uncontrollable power surge. The surge caused a sudden increase in heat that vaporized reactor cooling water and led to a steam explosion. This explosion exposed the core of the reactor to the environment. In the absence of an effective cooling mechanism, the core caught fire, releasing radiation into the atmosphere via smoke from the burning radioactive elements ("Chernobyl Accident and Its Consequences").

Operator "Error"

The International Nuclear Safety Advisory Group (INSAG) released an initial report on the Chernobyl accident (INSAG-1) that places the blame for the accident squarely upon operator error. Originally it was reported that "continuous operation below 700 MW(th) is forbidden by normal safety procedures owing to

problems of thermal-hydraulic instability" (INSAG-7). Under this assumption, the operators blatantly broke the established procedures and were to blame for the disaster. However, a later report, INSAG-7, reveals that nothing in the safety procedures forbade such operation. In other words, while the action ultimately proved disastrous, the operators did not violate operating procedures by running the plant under 700 MW(th). The operators did, however, knowingly initiate the test at 200 MW(th) even though the test procedure called for 700 MW(th) (INSAG-7). While technically not violating explicit policies, they purposely altered the test procedures to fit the current state of the reactor. This willingness to trivially change prescribed procedures was a symptom of a larger deficiency in safety culture within the operation of the plant. As the INSAG-7 report states: "Awareness of the necessity of avoiding such a situation should be second nature to any responsible operating staff and to any designers responsible for the elaboration of operating instructions for the plant." While the operators were a key component of the disaster, deeper analysis of the accident shows that system design, improper procedures, inadequately trained staff, and a deficient safety culture all played major contributory roles in the accident.

Flawed Design

The Chernobyl reactor was created in a manner that would fail if put into the wrong operating conditions. The problem of design is often a complex one. Uncertainty with how best to address forecasted issues along with how to predict where things might fail does not make the problem any easier. As the risk associated with system use increases, however, so does the responsibility of designers to understand and address any potential flaws in design. One major flaw of the RBMK reactor found within Soviet nuclear power plants like Chernobyl was the potential for operators to place the reactor in a state where it can fail. Directly addressing such a design flaw, the INSAG-1 report states the "Nuclear plant designs must be as far as possible invulnerable to operator error and to deliberate violation of safety procedure" (INSAG-7). In other words, the system should ideally be designed in such a manner as to prevent both intentional and unintentional human action from causing a disastrous result. This leads to the obvious lesson that it is important to design things well. While the design was indeed poor, the accident was still preventable.

Failure to Learn from Previous Mistakes

Indicators of design flaws with the Chernobyl plant were known years before the 1986 accident occurred. In 1975, in a similarly designed plant with a RBMK reactor at Leningrad, an accident occurred that "indicated major weaknesses in the characteristics and operation" of the reactor system (INSAG-7). Additionally, a 1983 incident at the Ignalina plant in the Lithuanian Republic alerted engineers to the deficiency in safety rods that factored into the Chernobyl disaster. The chief engineering organization for these reactors acknowledged this issue with the rods and promulgated the information to other plants; however, no changes were made to correct the problem (INSAG-7). Nuclear plants with the RBMK reactors were informed that restrictions would be made to prevent full withdrawal of control rods from the core, but these restrictions were not imposed and "apparently the matter was forgotten" (INSAG-7). The consensus view apparently was that the conditions that could lead to an accident would never occur, so taking action was unnecessary. These conditions that would never occur, however, just happened to do so in 1986 at Chernobyl.

The 1975 and 1983 accidents indicated the existence of specific design problems that contributed to the Chernobyl accident, but they were essentially ignored. This reveals a broader organizational failure at the nuclear regulatory level. The INSAG-7 report points to a "lack of communication and lack of exchange of information" as a reason for why the operating staff at Chernobyl were unaware of the design issues and consequences their actions would cause: a failure in the communication system within the broader Soviet nuclear regulatory system.

Organizational Deficiencies

Although the design of the RBMK reactors was clearly deficient, many opportunities presented themselves to prevent the accident at Chernobyl. While these opportunities existed, inadequacies of the safety culture of the Soviet nuclear system as a whole prevented them from being capitalized upon. Specifically, the foreshadowing incidents at Leningrad and Ignalina resulted in little modification to the reactor design or operating procedures. No independent safety body further analyzed the accidents, and the information about the accidents was not effectively passed on to the Chernobyl plant operators so that they could learn from the previous mistakes. Additionally, key safety operating procedures, such as forbidding operation under 700 MW(th) power, did not even exist. These issues indicate a lack of formalized communication pathways, and an even broader lack of appropriate regulatory structure and oversight. The deficiencies on the large scale prevented operators on the small scale from being appropriately educated and trained in safe operation of the poorly designed reactors. As INSAG-7 states: "It is reprehensible that such a deficiency had been known of for so long without its having been eliminated."

Individual Deficiencies

While the previously discussed issues rightly place a spotlight on the organizational deficiencies of the Soviet nuclear system, improper operator actions also point to individual deficiencies with the microcosm of Chernobyl. Blame for inadequate training of the reactor operators can reasonably be placed on the larger system, in that operators could not know what they were not told. What the operators did know, however, was that they should properly follow formal test procedures. Concerning the importance of test procedures: "Well planned procedures are very important when tests are to take place at a nuclear plant. These procedures should be strictly followed. Where in the process it is found that the initial procedures are defective or they will not work as planned, tests should cease while a carefully preplanned process is followed to evaluate any changes contemplated" (INSAG-7). Rather than stopping the test when the reactor power fell below 700 MW(th), however, the operators "did not stop and think, but on the spot they modified the test conditions to match their view at that moment of the prevailing conditions" (INSAG-7). These actions point to individual deficiencies within the small scale environment of Chernobyl that also contributed to the disaster.

3.2.3 Three Mile Island

Although labeled the "most serious accident in U.S. commercial nuclear power plant operating history," the partial reactor meltdown at Three Mile Island (TMI) in 1979 thankfully resulted in no fatalities or discernible negative health effects for workers or the public ("Three Mile Island Accident"). Much like the accident to occur a few years later in Chernobyl, TMI was the result of numerous organizational and operational failures. The accident led to many changes in U.S. nuclear regulation and operations, including greater emphasis in human factors engineering and reactor operator training.

The Accident

The 1979 accident at TMI involved a pressure relief valve that became stuck open after initially operating as designed. Reactor coolant began draining through the open valve, but the plant operators were unaware of this drainage and proceeded to misdiagnose the problem for more than two hours ("Lessons From the 1979 Accident at Three Mile Island"). Without the coolant, about half of the reactor fuel melted. In the end, a negligible amount of radioactive air escaped the plant while all the reactor fuel and radioactive cooling water were contained.

Lack of Human Factors Engineering

Similar to Chernobyl, both organizational and individual factors contributed to the accident. As the Kemeny report states, "The accident at Three Mile Island (TMI) occurred as a result of a series of human, institutional, and mechanical failures" (Kemeny et al., 1979). Operator "error" presents the most obvious explanation of the accident, specifically, the error of not correctly diagnosing and fixing the problem. The indicator light for the relief valve only indicated if a signal had been sent to close the valve; it did not show if the valve was actually closed (Kemeny et al., 1979). Relying on the indicator light, the TMI operators incorrectly assumed the valve was closed and did not correctly take into account indicators that suggested otherwise.

While the actions of the TMI operators were erroneous, they are understandable given the lack of human factors engineering within the control room design. Indicative of this issue, the Kemeny report states that the TMI control panel is "huge, with hundreds of alarms, and there are some key indicators placed in locations where the operators cannot see them." In general, the information in the control room was "not presented in a clear and sufficiently understandable form" (Kemeny et al., 1979). More than 100 alarms went off during the first few minutes of the accident, which led to confusion about what was most critical and should be attended to (Kemeny et al., 1979). This accident highlighted a lack of awareness for the importance of human factors engineering within the control room design and operation. According to a General Accounting Office (GAO) report on the matter, before TMI, human factors engineering was "virtually ignored" in control room design ("Three Mile Island: The Most Studied Nuclear Accident In History"). This GAO report claims that recognition of the need for including human factors engineering within the "design and operation of nuclear power-plants" emerged as one of the most critical lessons learned from the accident.

Improper Training and Procedures

Lack of human factors engineering was just one indication of deficiencies on the part of U.S. nuclear regulation. The Rogovin report on TMI harshly criticizes the Nuclear Regulatory Commission (NRC), saying that "the principal deficiencies in commercial reactor safety today are not hardware problems, they are management problems" (Rogovin et al., 1980). Improper training and established procedures ranked among some of those management problems. Operators did not have any written emergency procedures that addressed the issue of loss of coolant through a valve (Rogovin et al., 1980). Without established emergency procedures, the operators needed to rely upon their own problem solving ability, the failure of which highlights the inadequacy of their training. As the Rogovin report states, "the operators on duty had not received training adequate to ensure that they would be able to recognize and respond to a serious accident during the first hour or two after it occurred."

Failure to Learn

Similar to the Chernobyl disaster, the TMI accident was foreshadowed by smaller incidents at other nuclear power plants that were not communicated and learned from. In fact, in "virtually identical" scenarios, the TMI accident had almost occurred in two earlier incidents (Rogovin et al., 1980). The first occurred in 1974 at a reactor in Switzerland, and the second in 1977 at the David Besse plant in Ohio. Both of these incidents included the same failure of the pressure release valve coupled with misleading indicators concerning the amount of water within the coolant system. The difference between TMI and these instances, however, is that operators were able to correctly diagnose and solve the problem within minutes (Rogovin et al., 1980). The NRC did not know about the Switzerland incident until after the TMI accident; however, the NRC not only knew about, but also closely studied, the similar case at David Besse. This study of David Besse revealed the potential for the safety issue posed by a stuck open valve coupled with misleading indicators;

however, the results of this study were never communicated to the operators of TMI before the accident (Rogovin et al., 1980).

The Kemeny report provides a clear and insightful summary of the contributing factors to TMI, much of which also ring true for Chernobyl: "while the major factor that turned this incident into a serious accident was inappropriate operator action, many factors contributed to the action of the operators, such as deficiencies in their training, lack of clarity in their operating procedures, failure of organizations to learn the proper lessons from previous incidents, and deficiencies in the design of the control room."

Systemic Failures: Active and Latent

Rather than viewed through the lens of human operator failure, the accidents at Chernobyl and TMI can more appropriately be characterized as examples of systemic failure (Le Bot, 2004). In its description of the human factors behind adverse events, Reason (1995) points to two primary causes of systemic failure: active failures and latent failures. Active failures are the errors and violations caused by those who directly interact with the system. In the case of nuclear power plants, operator error would be considered active failure. In the case of systemic failures, however, active failures do not occur within a vacuum. Instead, they are often accompanied by latent, or hidden, errors made in the past, potentially years before any active failure occurs. These latent failures represent organizational decisions that carry consequences that may "lie dormant for a long time," and only become apparent when combined with other factors that lead to the ultimate failure (Reason, 1995). Organizational factors shape the environments that humans operate in, and these environments ultimately shape human behavior. Such a view of systemic failure does not absolve individuals from their contribution to active failures, but it also ensures appropriate examination of contributing organizational factors as well.

3.2.4 Cognition of reactor operators

The examples of Chernobyl and Three Mile Island primarily emphasize the importance of the organizational role in ensuring proper execution of decision-making within complex, model-centric environments. However, various cognitive challenges that nuclear power plant operators face also warrant consideration.

Monitoring, Complacency, and Vigilance

Most of a nuclear power plant's operations are automated, so a reactor operator's primary job is to monitor the functions of the power plant to ensure proper functionality and safety. Studies have demonstrated, however, that monitoring tasks can lead to a lack of vigilance and induced complacency (Molloy and Parasuraman, 1996; Endsley and Iris, 1995; Parasuraman, Molloy, and Singh, 1993). Complacency and decreased vigilance lead to a loss of operator situation awareness (SA) that may result in inappropriate decisions in the face of abnormal, and potentially dangerous, system behavior (Molloy and Parasuraman, 1996). Approaches such as dynamically varying task allocation between automation and operator control, and modifying the levels of automation used can potentially guard against the negative outcomes of complacency and vigilance decrement (Molloy and Parasuraman, 1996; Parasuraman, Mouloua, and Molloy, 1996). While the challenges of passive monitoring should very well be acknowledged and protected against when considering the role of the human operator within a nuclear power plant system, a discussion on monitoring may not prove the most relevant to this thesis' overall discussion of human decision-makers dynamically interacting with models in immersive environments. The role of the reactor operator is not passive, however.

Monitoring as Active Problem Solving

By investigating reactor operators in practice, Mumaw et al. (2000) reveals that "monitoring during normal operations was a complex, cognitively demanding task that was better characterized as active problem solving than as passive vigilance." This active problem solving involves the challenge of identifying and pursuing pertinent pieces of information against a "cognitively noisy background" (Mumaw et al., 2000). Characterizing "monitoring" as active problem solving of the plant through abstracted information in the control room lends itself more closely to the realm of decision-makers operating within immersive model-centric environments. Numerous factors make this problem solving within NPP control rooms challenging. For one, the system is very complex, consisting of thousands of components and instruments that operators gather information from. Exacerbating this complexity, "there are *always* components, instruments, or subsystems that are missing, broken, working imperfectly, or being worked on," due to the inevitability of failure when so many pieces exist within the system (Mumaw et al., 2000). Coupled with nuisance alarms, poorly designed displays and controls, and various types of automation, the task of monitoring and problem solving within a control room presents a cognitively demanding challenge (Mumaw et al., 2000).

Tailorability for Mental Model Development

While various factors make the monitoring and problem solving tasks of operators challenging, operators are able to interact with and structure their environment into one they can use to more effectively identify relevant information to guide decision-making. This tailorability presents a means for overcoming improper upfront design by allowing individuals to structure the environment to fit their preferences and contextual needs. These strategies compensate for human-system interface deficiencies and reduce cognitive demands through multiple means, including enhancing salience of important indicators and reducing background noise, creating new information, and offloading cognitive load through reminders and external aids (Mumaw et al., 2000). These strategies for effective monitoring are anything but passive, and reflect an awareness of the importance of developing and maintaining an accurate mental model of the system state in the operating and monitoring of a NPP.

Challenges to Mental Model Development

A report by the NRC titled "Cognitive Skill Training for Nuclear Power Plant Operational Decision Making," emphasizes the importance of operators developing an "accurate and complete mental representation" (or mental model) when monitoring the state of the plant (Mumaw, Swatzler, Roth, and Thomas, 1994). The report also notes specific challenges that can hinder proper mental model development. One such challenge includes over-relying on familiar patterns, also referred to as schemas. A schema is described "as a generic representation of a familiar or well learned event" and can be a useful tool for efficiently interpreting new events (Mumaw et al., 1994). The schema essentially provides a prefabricated mental model that can quickly be applied and tailored to more accurately understand the situation at hand. Highly familiar patterns or cues can sometime be misleading, however, and may prompt the application of incorrect schemas to a situation. When a new situation prevents operators from invoking a schema, inferences must be made in order to build a new mental model (Mumaw et al., 1994). Adding to the challenge of mental model development, there are limits on the amount of information humans can incorporate into a mental representation (Mumaw et al., 1994). These limitations are referred to as "representational limits." Finally, the report also lists confirmation bias as a challenge for effective mental model development. Decision-makers, knowingly or unknowingly, have expectations for the state of the system, and tend to seek information that confirms those expectations. This creates the risk that a mental model developed in this manner will be incomplete (Mumaw et al., 1994).

Appropriate Responsibility

While the accidents of Chernobyl and Three Mile Island emphasize the significant responsibility organizations and regulatory bodies have in setting operators up for success, analysis of NPP operators suggests the importance of providing operators with appropriate responsibility as well. No matter how much consideration for human factors goes into designing a control room interface, the designers will not be able anticipate all of the situations and decision-making scenarios that occur within the dynamically complex system of a nuclear power plant. Due to the limitations of designers, human operators must "continually finish the design (Mumaw et al., 2000). Granting operators appropriate latitude and responsibility, along with ability to tailor and "finish" the design, can achieve safer and more effective strategies for monitoring and problem solving than would be possible with complete rigidity and uniformity in operation.

3.2.5 Conclusions and Implications for Model-Centric Engineering

Examining the case study of reactor operators presents a complex picture of organizational, operational, and individual factors that intersect in the control room. The role of nuclear reactor operators, similar to a decision-maker in a model-centric environment, is one of active problem solving. This problem solving requires continuous creation of accurate mental models of the operating environment. The control room, much like a model-centric environment, should not be rigid, in that no upfront design will be appropriate for the wide range of possible situations one might encounter. Tailorability is a key trait of effectively operated NPP control rooms that not only allows the decision-making environment to be structured to help meet the needs of the moment, but also tailors the environment to the operator's preferences. Various challenges, such as poor schema selection, representational limits, and confirmation bias can hinder appropriate mental model development. Furthermore, this case study makes abundantly clear the influential impact that organizational factors can have upon the actions and decisions of individual decision-makers. Chernobyl and TMI illustrate the importance for having a strong organizational culture that strives for good upfront design, effective intra-organizational communication, proper training, intentional safety analysis, appropriate procedures, diligent enforcement of procedures, and the willingness and ability to learn from mistakes. These organizational and individual considerations gleaned from the experiences of NPP reactor operators may offer beneficial principles for guiding the future of model-centric engineering as well.

3.3 Decision-Making Theory and Bias Mitigation

The human mind is a remarkable thing: incredibly capable in many situations, but also prone to predictable errors in others. As this thesis aims hopes to help enable effective and efficient decision-making through model-centric engineering, consideration of individual judgment and decision-making warrants attention. This next section provides a discussion of dual-process theory, heuristics and biases, and bias mitigation.

3.3.1 Dual-Process Theory

Dual-process theory presents a model for thinking about human cognition, decision-making, and biases that plague our judgment. According to this theory, two systems, System 1 and System 2, govern our thinking and reasoning processes. Kahneman (2011) offers the following description of these systems:

- *System 1* operates automatically and quickly, with little or no effort and no sense of voluntary control. (Kahneman, 2011)
- *System 2* allocates attention to the effortful mental activities that demand it, including complex computations. The operations of System 2 are often associated with the subjective experience of agency, choice, and concentration. (Kahneman, 2011)

System 1, also referred to as the "automatic system," handles the automatic processes of attention and memory that let us operate efficiently and relatively effectively within much of our world. The primary purpose of System 1 is to "maintain and update a model of your personal world" (Kahneman, 2011). These models of familiar situations are often generally accurate and appropriate, as are System 1's initial reactions to many challenges.

System 2, on the other hand, handles slower and more effortful activities that require directed attention and conscious choice. These include processes such as higher reasoning and analytical problem solving, along with assessment of System 1's intuitive models, thoughts, and emotions. System 1 is responsible for causing you to notice and automatically jump out of the way of a speeding bicycle on the sidewalk, as well as for the feeling of annoyance that may subsequently arise. System 2, for its part, is the actor responsible for telling you to take a deep breath and restrain yourself from blurting out certain choice words at the passing bicyclist. Certain automatic functions, like the ability to discern emotion of anger on another's face or to perceive that an object is further away than another, are inherent to most humans. Other functions, like reading and comprehending a familiar word or solving a simple math problem, can be learned and acquired through directed use of System 2 processes. Kahneman (2011) uses math problems to illustrate one difference between System 1 and System 2 processing. For example, solve the following math problem:

$$2 + 2 = ?$$

For most of us, the answer "4" instantly pops into our head. Although this answer is not naturally intuitive, it has been programmed into our System 1 through effortful concentration and practice from our System 2 years earlier in grade school. Now, stop and solve this math problem:

$$16 \times 23 = ?$$

For the majority of us, solving this problem requires slower, concentrated attention to multiply and confirm the answer in our head, or to search out a calculator and type in the numbers to find the answer. In either case, directed attention was required to determine the correct answer (which is 368).

Reading offers another example of the differences between System 1 and 2 processing. For a sufficiently educated English speaker, the words on this page can be read relatively effortlessly using System 1

processing – essentially, we can read and comprehend words without thinking about it. For new students of English, or individuals learning to read, however, reading this page would require the concentrated and effortful attention of System 2 to process the words and achieve comprehension.

A Coherence-Seeking System 1 and a Lazy System 2

System 1 can be described as coherence-seeking, often jumping to conclusions to efficiently create coherent stories from whatever information has been received. System 2 is capable of governing and directing assessment of these conclusions and coherent stories, thereby preventing acceptance of improperly assumed conclusions. Much as flowing water follows the path of least resistance, however, System 2 often takes the path of least cognitive effort and may not properly monitor and catch erroneous "coherent" stories created by System 1. In short, System 2 is predisposed to laziness and will often "endorse many intuitive beliefs, which closely reflect the impression generated by System 1" (Kahneman, 2011).

If you can raed all the wrods in tihs setnence, even though varoius lettres are jubmled up, you msut be raelly smrat. If you are able to read the previous statement, while it unfortunately does not necessarily mean you are "raelly smrat" in that almost everyone else can read it too, it does provide insight into our brain's ability to create coherence from imperfect information. Perhaps the initial read through or two of this paragraphs first sentence required a combination of System 1 and 2 processing; however, after realizing the gimmick, you were likely able to read the sentence without significant impediment or concentration. This demonstrates one simple example of our System 1's ability to automatically create coherent pictures out of incomplete, or imperfect, information. As we can never have perfect knowledge, any understanding of a situation involves abstracting the imperfect information into a coherent mental model that guides our decisions. While the effortful and thoughtful processes of System 2 can guide more accurate mental model formation and subsequent decisions, the initial framework for the mental model seems to be a product of System 1's initial, automatic functions (Kahneman, 2011).

Gilbert (1991) puts forth the theory that the comprehension of a statement or argument must first begin with belief – we must first accept the statement as true before we can truly understand what it means. Following this comprehension that requires initial belief, we can then assess and deconstruct the statement in order to "unaccept," or at the very least suspend in the absence of further information, the belief (Gilbert, 1991). Gilbert goes on to state, however, that "[p]eople are credulous creatures who find it very easy to believe and very difficult to doubt" (Gilbert, 1991). This unraveling of the momentary belief required for comprehension takes effort and time, functions handled by System 2. The dual-process theory, combined with Gilbert's theory of comprehension, if taken to be true (and can we not help but initially take it to be true?), presents concerning implications. As stated by Kahneman (2011): "System 1 is gullible and biased to believe, System 2 is in charge of doubting and unbelieving, but System 2 is sometimes busy, and often lazy."

3.3.2 Heuristics and Biases

In addition to automatically forming coherent mental models of situations based upon available information, System 1 also tends to "help" System 2 answer complex problems by generating and solving tangentially related, but much simpler, questions. Kahneman (2011) defines answering one question in place of another as "substitution," the "heuristic question" as the simpler question substituted in place of the actual "target question." A heuristic, within this line of thought, is defined as "a simple procedure that helps find adequate, though often imperfect, answers to difficult questions" (Kahneman, 2011). For example, before sitting in a chair, you might want to know if the chair will support your weight. You could thoroughly inspect the chair for quality or look for a label indicating maximum allowable weight; however,

the question of the chair's adequacy could be substituted with other questions that are simpler to answer, such as: have similar chairs failed me before, or, do I trust the chair's manufacturer?

In fact, much of our society's systems of regulations, law, and social norms have evolved to allow us the efficiency of using heuristic judgments, such as the evaluation of trust, in place of more specific questions. You can trust the chair will reliably hold your weight because the chair manufacturer may face liability charges if it does not, or at the very least it will not last long in a competitive market. Within the societal environment we live, therefore, not only is it rational to apply heuristics, but within many reasoning and decision-making contexts it would be untenable *not* to do so.

As discussed, System 1, with its jumping to conclusions, coherence building, and heuristic substitution, presents a means for generating intuitive opinions on complex issues (Kahneman, 2011). If substitution occurs, System 2 has the opportunity to endorse, reject, or modify the intuitive conclusions presented to it. However, as mentioned before, System 2 can be lazy, and "often follows the path of least effort and endorses a heuristic answer without much scrutiny of whether it is truly appropriate" (Kahneman, 2011). So herein lies the problem: heuristic generalizations effectively offer efficiency in navigating many situations in life, but they also fail in specific contexts, leading to predictable biases and errors in judgment. In many situations the consequences of errors in judgment may be small enough to not warrant vigilant assessment of decisions. In others, however, as the glass cockpit and reactor operator case studies have shown, biases and judgmental errors may be unacceptable due to the high level of risk they pose.

Biases represent "predictable deviations from rationality," of which over a hundred specific biases are known (Croskerry, Singhal, and Mamede, 2013a). Central to this definition is rationality; if we do not know what a "rational" decision is, we cannot rightly say when a deviation occurs. Rationality is comprised of two components: correct beliefs and correct decisions. Stanovich, West, and Toplak (2012) offers the following view on rationality: "For our beliefs to be rational they must correspond to the way the world is – they must be true. For our actions to be rational, they must be the best means towards our goals – they must be the best things to do." Due to the nature of complex problems and limited information, however, it is often impossible to discern what is fully true and what is the absolute best thing to do. In this case there exist degrees of rationality, in that certain decisions may be more rational than others. While the best possible decisions may be challenging to identify, poor decisions, or deviations from rationality, are much easier to spot (Stanovich, 2011). Biases represent these deviations that comprise identifiable and irrational behavior.

Examples of Biases

While over a hundred biases have been identified in literature (Croskerry et al., 2013a), below are a few examples of biases that commonly affect our judgment:

- Automation bias
    - As noted in the glass cockpit case study, Mosier and Skitka (1999) define automation bias as "the use of automation as a heuristic replacement for vigilant information seeking and processing." This bias can result in commission errors (incorrectly following an unverified automation directive) and omission errors (failing to identify an issue not identified by an autonomous system).
- Confirmation bias
    - Confirmation bias leads people to unwittingly seek or interpret evidence "in ways that are partial to existing beliefs" (Nickerson, 1998). This bias "leads people to ignore evidence that contradicts their preconceived notions," (Kahneman, Lovalla, and Sibony, 2011) and

rather, to utilize evidence that "confirms" their positions in a "one-sided case-building process" (Nickerson, 1998).

- Anchoring bias
    - Anchoring bias leads individuals to "weigh one piece of information too heavily in making decisions" (Kahneman, et al., 2011). Oftentimes, people make estimates from an initial starting value that is then adjusted to arrive at the final estimation. These estimates are often strongly biased towards staying near the starting value, in other words, they are "anchored" to the initial value (Kahneman, 2011).
- Availability bias
    - Availability bias leads individuals to judge how often an event occurs "by the ease with which instances or occurrences can be brought to mind (Tversky and Kahneman, 1973). Essentially, easily recallable situations will be estimated to occur more frequently than situations that are harder to recall. This might lead to someone overestimating the likelihood of a commercial airline crash because they saw a headline of a crash in a recent newspaper, when in fact, the probability of an accident is quite low.

3.3.3 Bias Mitigation

Bias clearly can have detrimental effects in many high-risk fields, from medicine, to law, to business – with literature published on bias mitigation within each of these respective fields (Croskerry et al., 2013a; Burke, 2007; Hammond, Keeney, and Raiffa, 1998). The insidious thing about System 1 processing is that we are often unaware of its influences and cannot ever turn it off (Kahneman, 2011). Psychologists know how to describe biases and how people should normatively act, but less is known about how to help people mitigate these biases and behave optimally (Milkman, Chugh, and Bazerman, 2008). Some claim that "forewarned is forearmed," and that "the best defense [against biases] is always awareness" (Hammond et al., 1998). However, while awareness may be a crucial component of any type of personal mitigation, it may not prove effective in and of itself. So while we might know we are prone to biases, we are poorly predisposed to prevent them on our own. Activation of System 2 processes to counteract improper System 1 heuristics and conclusions is ultimately the key to suppressing and mitigating biases; however, as noted, we are often blind to our own biases, and it would prove impossible to constantly question all of our decisions (Kahneman, 2011). To be clear, System 2 processing is still vulnerable to conscious mistakes and errors, but the focus here is not on conscious errors of judgment, but rather the unconscious. Kahneman (2011) suggests that the best comprise is learning to identify and mitigate situations where biases are more likely and the risk is high. However, Kahneman (2011) also claims that "it is easier to recognize other people's mistakes than our own."

Literature on cognitive debiasing suggests two primary pathways for bias mitigation within an organization: strategies that aim to prevent biases by improving individual decision-making, and strategies that accept the inevitability of biases but set up pathways, or "forcing functions," to catch biased decisions and to prevent situations in which biases may occur (Croskerry, Singhal, and Mamede, 2013b). Strategies aimed at improving individual decision-making fall under the general categories of education and training. This form of mitigation includes developing an awareness for biases and how they occur. "Forewarned is forearmed," and if decision-makers can be appropriately forewarned about situations where biases may occur, even if they are personally unaware of them, they may develop triggers for consciously moving to System 2 processing. For example, if decision-makers find themselves getting excited about a proposal that seems "too good to be true," this could trigger a recognition that they may need to be more skeptical. This could then force them to consider alternative explanations to prevent a confirmation bias where they only seek out supporting evidence. In order for individuals to mitigate personal bias, they must be aware of the

potential for biases, and then be able to detect and appropriately respond by suppressing biases in situations where they may be likely to occur (Burke, 2007). Education and training allow people to develop this awareness for biases, along with strategies to prevent them.

According to Burke (2007), however, while some empirical evidence suggests that bias can potentially be mitigated through education, other evidence suggests "education is an unlikely panacea." Kahneman et al. (2011) more explicitly argues that education is not effective at eliminating biases, claiming that "You may accept that you have biases, but you cannot eliminate them in yourself." Bias mitigation is not without hope, but rather than focusing on the individual, the focus should be at the collective, organizational level (Stanovich, West, and Toplak, 2014). Analysis of relevant literature suggests that this organizational mitigation has two potential strategies: (1) creating forcing functions aimed at making individuals use System 2 processing in place of System 1 in bias-prone situations, and (2) developing pathways for recognizing bias in others. These forcing functions could manifest as organizational policies, or rules, that require System 2 activation in certain circumstances. An example would be requiring individuals to follow checklists that prescribe predetermined "rational" behavior for given situations (Kahneman et al., 2011). This is well-demonstrated by aircraft pilots who use checklists in a manner that forces System 2 processing to analytically follow prescribed steps. Another example could be a standing policy that requires assigning a "devil's advocate" to provide counter-arguments to prevent confirmation bias or group-think (Hammond et al., 1998).

While we might not be very good at recognizing our own biases, Kahneman et al. (2011) argues that "we can apply rational thought to detect others' faulty intuition and improve their judgment." As most decision-making processes involve multiple individuals, the second organizational strategy can focus on enabling pathways to detect and prevent biases in others. One such means could include the use of non-advocate reviews that are conducted through internal or external processes. Creating policies that require review of decisions enables experienced individuals to catch and mitigate errors of judgment in others. This is where education and training also proves to be crucial. If reviewers are unaware of the potential for biases in others' decisions, they may be as blind to the hidden biases as those who committed them. Kahneman et al. (2011) proposes a mix of forcing functions and review-based organizational mitigation, namely, a checklist of revealing questions for reviewers and final decision-makers to ask.

This checklist is aimed at supporting executive decision-makers in identifying potential biases present in business proposals set before them for review. While model-informed decisions may not strictly involve business decisions, the same principles for reviewing proposals seem relevant to "proposals" offered by models or individuals working with models. In each case, decision-makers need to review the information set before them and decided whether to accept the results, or if flaws earlier in the decision process make the results flawed or at least worthy of further review. The following section offers questions adapted from Kahneman et al. (2011).

Questions for identifying and mitigating bias (adapted from Kahneman et al., 2011)

1. Is there any reason to suspect motivated errors, or errors driven by the self-interest of the recommending team?

Answering "yes" to this question should alert a decision-maker that those presenting a recommendation may have a higher potential for unintentional self-deception and biased judgment. People tend to be biased "in the direction of their own interests" (Kahneman et al., 2011).

2. Have the people making the recommendation fallen in love with it?

This question aims at uncovering the influence of the affect heuristic – making judgments based on how we feel about something. People are biased to overestimate benefits and minimize the costs of things we like, and to do the opposite when assessing things we do not like (Kahneman et al., 2011).

3. Were there dissenting opinions within the recommending team?

Groupthink occurs when people in a decision-making group seek to avoid conflict and end up favoring decisions that generate group support, even if these decisions are biased (Kahneman et al., 2011). Dissenting opinions can help uncover improper assumptions and judgments. Groupthink, however, may prevent a healthy form of discussion and dissent.

4. Could the diagnosis of the situation be overly influenced by salient analogies?

Analogies, like models, offer value in certain situations, but also have limitations. Relying too heavily on analogies as arguments may lead to faulty assumptions. The availability bias may also lead individuals to place undue weight upon easily recallable, or salient, analogies, while discounting other important factors or examples that do not come as readily to mind (Kahneman et al., 2011).

5. Have credible alternatives been considered?

Confirmation bias can lead people to generate a single hypothesis, and then to only seek "evidence that supports it" (Kahneman et al., 2011). Forcing individuals to consider the pros and cons of multiple alternatives may help uncover relevant information that was not previously found through information seeking influenced by confirmation bias.

6. If you had to make this decision again in a year, what information would you want, and can you get more of it now?

The coherence building tendency of our intuition can lead us to develop cohesive understandings of situations that might not acknowledge holes in data and judgment. This question forces decision-makers to consider what else could be useful – potentially uncovering additional information that could aid in a more effective decision.

7. Do you know where the numbers came from?

This question helps to counteract anchoring bias – where decisions and estimations are unduly influenced by initial numbers from which estimates are then adjusted. Sometimes initial numbers may only be best guesses rather than facts, however, these numbers can create a bias that anchors adjustments near to these guesses (Kahneman et al., 2011).

8. Can you see a halo effect?

The halo effect biases people to view a situation "as simpler and more emotionally coherent than it really is" (Kahneman et al., 2011). This effect can lead individuals to assume that everything about an individual or organization is exemplary because those entities have been labeled "excellent" by other experts. They may very well be excellent in certain matters, but that does not mean they are excellent in all matters. Additionally, "excellence" in certain outcomes may be more due to luck and timing rather than skill, and may not lead to the same results in the future (Kahneman et al., 2011).

9. Are the people making the recommendation overly attached to past decisions?

When money is spent it is gone, sunk, and should not have a bearing on future decisions. The sunk-cost fallacy, however, can lead individuals to let present decisions about the future be inappropriately influenced by past decisions and expenditures in an effort to justify past, flawed choices (Kahneman et al., 2011; Hammond et al., 1998).

10. Is the base case overly optimistic?

Individuals can be influenced by an overconfidence effect that leads them to place too great of weight or optimism on their decisions. This tendency can be especially prevalent amongst groups or individuals that have had a recent track record of success (Kahneman et al., 2011).

These questions were originally designed to interrogate individuals who are presenting recommendations to a decision-maker. People are poor at recognizing their own irrational biases, so these questions offer a tool for a decision-maker to help reveal hidden biases in others' judgments. Within a model-centric engineering context, decision-makers are not so much directly interfacing with individuals, but with models. These models, however, are products of real people who are also just as susceptible to biases. These questions, therefore, whether being targeted at recommendations from people or from a model, can help uncover biases that may unwittingly be hidden beneath the shiny surface.

3.3.4 Conclusion

Science will likely never fully understand the complexities of the human mind. The System 1 and System 2 conceptualizations of dual-process theory are simply abstractions geared towards aiding a better understanding. They are models. Such conceptualizations are subject to the limitations and inaccuracies inherent within all models, and as such, are not without valid criticisms. For example, although Stanovich (1999) first suggested the terms "System" 1 and 2, Stanovich has since switched to using the terms Type 1 and Type 2 to avoid suggesting that there are two distinct brain systems responsible for these observed behaviors (Stanovich et al., 2014). Although limited as it may be in predictive and descriptive power, the model of System 1 and System 2 can prove valuable for describing and understanding how and why humans think and act as they do. While this understanding can prove insightful in and of itself, its true value stems from what we can do with this information. For the purposes of this thesis, this discussion aims to contribute to a more holistic understanding of decision-making within a model-centric context, with hopes that this understanding may also lead to strengthened model-centric decision-making.

3.4 Summary of Model-Centric Challenges and Mitigations

Model-centric engineering seeks to enable more efficient and effective technological decisions. While this chapter does not claim to offer insight into all the potential challenges that may hinder effective model-centric decision-making, it does hope to present various specific challenges and ideas that are worth considering in the development and use of model-centric practices and policy.

Glass Cockpit Case Study

The glass cockpit case study presents various specific challenges that occur when pilots interact with the complexities of automated and improperly transparent cockpit systems. Pilots must make decisions based on the abstracted aircraft information presented to them through the glass cockpit displays; however, cognitive and perceptual challenges can hinder appropriate understanding and pilot response. Automation can lead to an automation bias and automation-induced complacency that increases the likelihood of pilots making mistakes through acts of commission and omission. Additionally, while modes offer a means for

tailoring automation and system capabilities to pilot preference and situational context, improper transparency can lead to mode error. This mode error manifests itself through automation surprises, where the aircraft operates in an unexpected manner because pilots are unaware of the mode it is actually operating in. Assigning clear accountability helps prevent the negative effects of automation bias, and proper transparency aids pilots in understanding the current mode of operation. This case study highlights the fact that automation does not displace the role of humans within a system; rather, it changes it, and these changes can have both positive and negative consequences. With such changes we should not expect humans to perfectly adapt to their new roles, but should rather design the changes with the humans in mind and adapt the system to meet the needs of those ultimately responsible – the human operators. As models increase in integration and complexity, automation may play a role in managing and tailoring the abstracted information presented to model-centric actors and decision-makers. While not a complete encapsulation of all the challenges that automation may present to model-centric decision-making, this case study highlights specific potential challenges, along with the broader message emphasizing the need for considering the human within system design.

Nuclear Reactor Operator Case Study

Nuclear reactor operators offer another useful empirical case study for analyzing individuals within complex, abstracted decision-making environments. The disasters of Chernobyl and Three Mile Island illustrate an understanding that failures should be viewed through multiples lenses –rightly considering the improper actions of individuals that directly led to the failure, but also the broader organizational factors that share a large responsibility for the failure as well. Organizations have the ability and responsibility to shape the physical, procedural, and regulatory environments that individuals and decision-makers operate within. These environments significantly affect what the operators can and cannot do, along with how capable they are to perform at the level expected of them. Initial system design should explicitly take these human factors in account; however, due to the impossibility of predicting all potential futures, the system will inevitably contain flaws. Proper review procedures should be coupled with established communication pathways to discover and share relevant design flaws so they can be fixed and the operators appropriately trained to compensate for them. This suggests a need for upfront acknowledgement of the inevitability of design flaws, and a predetermined commitment to identifying, addressing, and appropriately empowering operators to iteratively build a more effective overall system. The acknowledgment of the significance of organizational factors may prove an important consideration for the effective design and use of model-centric environments in the future.

Decision-Making Theory

Dual-process theory offers a model for breaking human thinking and judgment into two systems: System 1 and System 2. System 1 operates autonomously and effortlessly, guiding our intuitive impressions, emotions, and answers to specific problems. System 2, on the other hand, operates more slowly and effortfully, and is responsible for reasoning and analytical processes that require conscious allocation of attentional resources. System 1's ability to quickly jump to conclusions by applying generalizations, or heuristics, to situations allows us to efficiently and often effectively make decisions throughout many day-to-day activities. These heuristics are effective in many cases, but are limited in scope and may fail to guide rational action in other situations. This leads to predictable deviations of rationality known as biases. System 2 is ultimately in charge of final judgments and decisions; however, the systematic biases within System 1 can influence System 2 processing and ultimately lead to irrational decisions. While conscious use of System 2 to monitor and suppress improper System 1 heuristics is needed to prevent biased decision-making, we are often unaware of our own biases, making personal bias mitigation challenging. Organizational structure can "force," or encourage, certain practices and behaviors that prove effective in

preventing and identifying biases within decision-making processes. Checklists are an example of practices that force System 2 use to prevent improper intuitive judgment in bias-prone situations. As another example, non-advocate reviews offer individuals the opportunity to uncover biases within others' decisions. An awareness of potential biases and pitfalls is crucial to identifying biases within oneself and others, and asking specific questions may help uncover hidden biases that were previously unknown. This section on dual-process theory and biases concludes with a list of questions that may be useful for mitigating biases within model-centric decision-making.

4 Expert Interview Study

This chapter explores various dimensions of enabling model-informed decisions, as motivated by the increasing need for individuals and teams to make decisions based on models and model-generated information. Central to this topic is the need to understand what engenders trust in models. This exploratory study uses expert interviews to investigate how various types of decision-makers and actors interact with and use models, including to what degree models are used to inform system decisions and how individuals build trust in models. While anecdotal stories of success and failure exist, empirical studies are needed to truly understand the many facets of human decision-making in model-centric engineering. Such research is expected to generate key insights that can inform current and future practice, as well as determine areas for more extensive study.

4.1 Introduction

This study aims to generate insight into decision-maker trust and perception of models and model-generated information. Experts in system decision-making accumulate various kinds of knowledge and wisdom, often through years of hard-earned experience. Rather than theorize on how various actors interact with and trust models, an interview-based approach allows us to gather qualitative, empirical data by asking them directly. This study is not meant to offer definitive truth for all types of decision-making with models, but rather to serve as an exploratory study into a little-researched area. This study is primarily scoped to decision-makers and systems experts found within the defense and aerospace communities.

Sampling

Unlike quantitative research, which advocates random sampling approaches, qualitative research seeks to "select 'information-rich' respondents who will provide you with the information you need" (Kumar, 2011). For this study, we have primarily used judgmental and expert sampling to identify "persons with demonstrated or known expertise in an area of interest," (Kumar, 2011) along with individuals who, although perhaps not widely known as "experts," were judged to have experience relevant for achieving the objectives of the study. In this study, we broadly view an expert as an individual who works or has worked as an actor within model-based decision-making processes, and can provide knowledgeable insight and perspective informed through his or her experiences. The definition of an expert is clearly open to interpretation as an "expert" may very well be in the eye of the beholder, and an improper interpretation on the part of researchers may lead to a biased sample of participants that fails to adequately represent a population. From our perspective, however, all participants were judged to have relevant experience and credentials through either their individually known work with models or that of the organizations for which they have worked, primarily experience found within the domains of defense and aerospace. While the study is ongoing, thirty individuals have been interviewed at the time of writing.

Interviews for this study followed a semi-structured format that allowed interview participants latitude to share a wide range of perspectives and insights while following guiding questions aimed at generating insight for the study objectives. Table 1 presents a list of the general questions asked.

Table 1. List of Interview Questions

1. What types of decisions do you make, or help others make, with models?
2. What is the degree to which the decisions you make are based on models?
3. Do you view models as a primary or supplementary source in decision-making?
4. How do you develop trust in models?
5. How do you judge if a model can be trusted?
6. How much transparency do you desire?
7. What factors have led to inappropriate trust in models?
8. What limits your ability to use models to make decisions?
9. What challenges or failures have you experienced with the use of models in system decision-making?
10. What approaches or policies have been applied, or would you like to see applied, to mitigate those challenges?
11. How desirable would the ability be to directly interact with models real-time while making decisions?

4.2 Decision-Making Flow of Model-Generated Information

High-level decisions incorporating an explicit model in the decision-making process include the following broad components:

1. A model that represents some aspect of the system of interest
2. Human actors
3. A decision to be made

While simplified, a generic conceptualization of the model-influenced decision-making process is helpful for facilitating discussion surrounding this research space. In this general framework, the information generated from a model is the common thread that connects the three generic commonalities listed above. First, a model must be created or already exist before it can be of use in a decision-making context – this creation of the model itself generates information relevant to decision-making before it is even "used." Next, the model in question generates information designed to facilitate a better understanding of an issue for which a decision must be made. Exactly what happens to this model information varies from context to context, but all contexts involve model information flowing *from* an actor (i.e. modelers or analysts directly interacting with the model), *through* another actor or actors, and ultimately *to* a final decision-making actor (Figure 6). Where decision-makers reside in the process seems to be more along a spectrum of the flow of model information. Within different decision-making contexts, actors may even find themselves in different roles. For example, in a mid-level decision-making context, an actor may be the individual *to* whom the information is flowing, yet in a higher-level decision, the same individual may become a *through* actor. In decisions involving more than one actor, however, all model-informed decisions involve information being generated and flowing from those directly interacting with a model, flowing through an actor or actors, and lastly reaching a final decision-maker to whom the information flows.

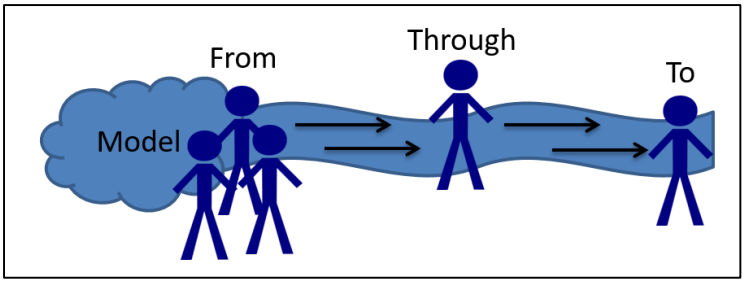


Figure 6. From-Through-To Flow of Model-Generated Information

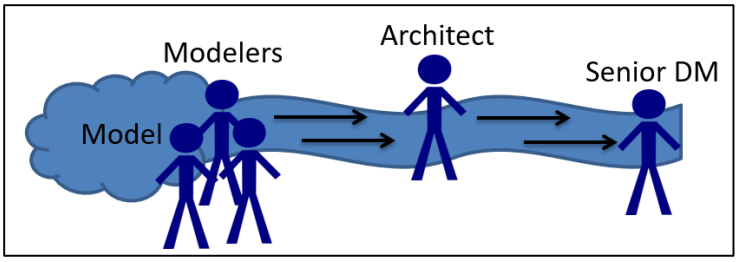


Figure 7. Modeler-Architect-Senior DM Flow

To elucidate this conceptualization, it may be helpful to examine a specific case example that illustrates this flow of information. Figure 7 illustrates one such scenario where a model is used to inform decision-makers in a war game. The senior level decision-maker (Senior DM) identifies a modeling need, and interfaces with a model architect to create the desired model. The architect works with a team of modelers who develop and test the model and produce model outputs that are communicated through the architect to the decision-maker in response to specific queries. In this case, the model information flows *from* the team of modelers who comprise the initial actors, then flows *through* the primary model architect, and finally *to* the Senior DM involved with the war game.

At the end of the model-generated information flow there is a fairly discrete decision or set of decisions to be made. These high-level decisions, however, are influenced by countless smaller decisions and actions performed by various individuals within the flow of information. This study seeks to better understand the perspectives and thought processes of these various actors with the hope of better understanding the decision-making process as a whole. In the sampling process we sought perspectives from individuals from all three of our conceptualized categories; however, for the purposes of this paper, *through* individuals comprise the majority of participants. While this study is ongoing, we believe the thirty individuals interviewed at the time of writing present enough information to warrant publishing of the current results.

4.3 Trust

Ricci, Schaffner, Ross, Rhodes, and Fitzgerald (2014) describes how trust in models relates to a user's perception of how close to a specified reality a constructed model is perceived to be. Ultimately, a good decision "is one based on a trusted, truthful representation of both reality and values" (Ricci et al., 2014). The 2015 IMCSE Pathfinder Workshop report (Rhodes and Ross, 2015) notes that numerous challenges exist within model-centric development, including challenges surrounding "perception of truthfulness and trust" in models, as this aspect of trust can ultimately affect "the timeliness, quality, and confidence in model-based decisions." The Pathfinder report also expresses a desire not just for models to be trusted, but for that trust to be supported with underlying evidence. Blackburn, Bone, and Witus, (2015) articulates a vision for developing model-centric environments into a "single source of technical truth" for decision-makers. West and Pyster (2015) communicates the idea of digital system models offering an "authoritative representation" of systems. Gass and Joel (1981) notes, however, that all models "reflect modelers' views of how the decision problem can be resolved," and that these views carry inherent assumptions and limitations that decision-makers must consider prior to determining if the subsequent modeling results appropriately align with their decision at hand. With this in mind, the goals of developing single sources of "truth" and "authoritative data" will require decision-makers to evaluate and determine how much trust they should place in this data. This trust can be improperly calibrated, however, potentially resulting in overreliance or underutilization. Engendering an appropriate level of model trust within decision-makers is crucial to effective use of models in decision-making.

Literature addressing human trust in automation offers insight that can be useful when applied to this discussion on human trust in models. This relationship seems rather natural when considering that automation may arguably be nothing more than a model of operation algorithmically programed into a machine. The article "Humans and Automation: Use, Misuse, Disuse, Abuse" (Parasuraman and Riley, 1997) highlights multiple potential pitfalls to consider when placing humans into interaction with automation. Misuse is defined "as overreliance on automation (e.g. using it when it should not be used, failing to monitor it effectively), disuse as underutilization of automation, […] and abuse as inappropriate application of automation by designers or managers" (Parasuraman and Riley, 1997). While examining factors that may contribute towards use and application of automation, Parasuraman and Riley (1997) notes that "trust often determines automation usage." This taxonomy of use, misuse, disuse, and abuse can provide a useful framework for thinking about how humans interact with complex models as well.

But what exactly is meant by "trust?" Lee and See (2004) defines trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." Specifically addressing misuse and disuse, Lee and See (2004) expresses that "[o]vertrust is poor calibration in which trust exceeds system capabilities; with distrust, trust falls short of the automation's capabilities." This idea of calibration "refers to the correspondence between a person's trust in the automation and the automation's capabilities" (Lee and See, 2004). Trust in automation implies belief that the automation will do what it is supposed to do, while trust in models assumes that the models will provide the information you want. Both automation and models represent technologies that require a certain amount of trust as the underlying processes and assumptions may be difficult to fully understand. The goal is not just for models to be used, but to be used appropriately; models, much like automation, have limitations of effectiveness and applicability. Overreliance in models can lead to misuse by inappropriately applying models outside of their inherent limitations. Conversely, improper lack of trust in models can lead to decision-makers discounting relevant model information that could have otherwise aided in the understanding and solution of issues. By examining the human aspect of human-model interaction, this study aims to generate understanding that can lead to appropriate "calibration" of human trust in models. Before seeking to influence the human actors, however, it is necessary to understand how those actors actually work in practice.

Developing Trust

Consciously or not, decision-makers must have a certain amount of trust present before model-generated information is used in the decision-making process. Few of the actors interviewed have consistent processes to develop trust in new models, yet all have various factors they consider when determining trust. Some factors prove unique to specific individuals or groups of individuals along the flow of information, while other factors appear to be common for individuals throughout the entire flow. In addition to processes or factors influencing decision-maker trust, we want to know more about what specific attributes or types of information about models that decision-makers and actors care about knowing.

4.4 Key Findings

This section presents preliminary key findings of the study to date. While these findings may not necessarily be novel, they serve to form a compilation of empirical evidence concerning human-model interaction. As this work is ongoing, these findings are expected to grow and evolve. The results are presented in no particular order of importance.

Technological and Social Factors Influencing Trust

As summarized by one participant, "trust is terribly important" within the modeling and decision-making process. While few of the interviewed experts have a specific process used in determining trust, every participant has various factors that they consider while determining the amount of trust to put in a model. This trust is also very contextually dependent, meaning that the trust is not so much in the model as an entity, but in the usefulness of the model for a specific decision at hand. Various factors influence individuals' trust in models, yet these factors may vary in importance depending on the specific individual involved. A clear theme that has emerged from the interviews, however, is that both technological and social factors come into play when determining the amount of trust that any type of actor is willing to place in a model. In many cases, the importance of technological factors appears to diminish in relation to social factors as actors move further along the flow of model information. Figure 8 illustrates the concept that various technological and social factors influence a decision-maker's trust in a model. The factors listed are not all-inclusive, but represent some of the factors identified through the interviews. While there may be trends in comparing important factors between the "*from, through, to"* categorizations of actors, such as the generalization that social factors seem more salient for *to* actors than for *from* actors, this is still dependent on the specific individuals involved. A strongly supported generalization, however, is that both technological and social factors play an important role in influencing an individual's trust, and any attempt to understand trust without considering both types of factors would be lacking.

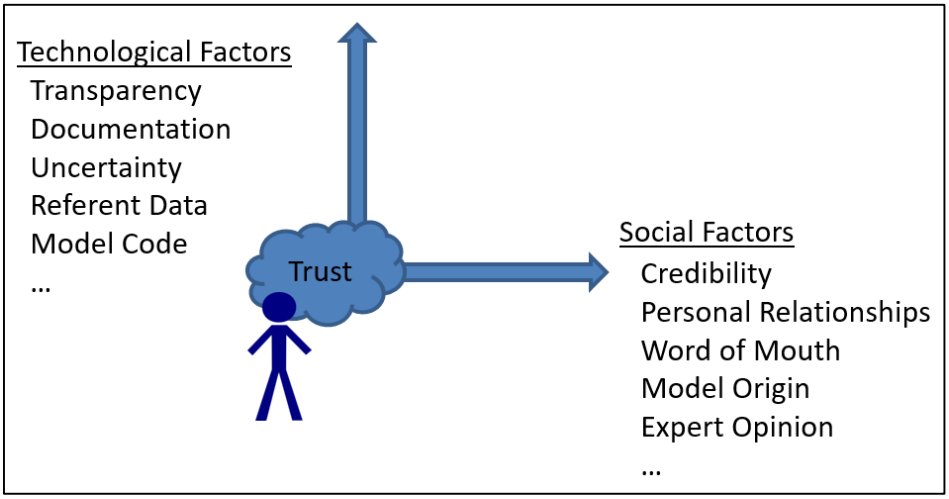


Figure 8. Sociotechnical Factors Influencing Model Trust

Importance of Communication

Communication arose as a key attribute of effective model decision-making. Before any effective modeling can be accomplished, senior level decision-makers must construct the problem statement clearly and in a form that unambiguously expresses the information they desire from models. Oftentimes the problem can change, however, therefore consistent communication of the problem at hand is crucial for allowing individuals below them to create or use models to generate relevant and useful information. The onus for this specific communication does not fall solely on senior levels, however, and lower levels must actively update senior decision-makers on progress to gain feedback on whether they are addressing the actual problem. Senior decision-makers must likewise be open and available, to the extent possible, to provide this feedback as necessary. As noted by an interview subject, "models […] bring their own language with them" that can create communication barriers that stifle decision-makers' understanding of the model output presented to them. Unless a decision-maker is similarly an expert in the model, there needs to be a

"translation between output to decision-maker speech," before the information can usefully be incorporated by the final decision-maker. Modeling aims to provide an asset for a decision; however, this asset cannot be effective if it is not useful for a decision-maker, and it cannot be useful if not understood. Instead of relegating discussion between actors to the beginning and end of a decision-making process, employing continuous and iterative communication may further reduce the acceptance barrier by allowing decision-makers to feel as if they walked the up-stream actors to the final model outputs. The flow of information between actors, including both expression and interpretation of information, must be intentional and unambiguous.

Transparency

Most of the interviewed modelers, analysts, and architects emphasized the importance of having access to precise technical information of models, oftentimes stating a desire to have access to code and the "guts" of the model in question. One such practitioner expressed that he "hopes everyone wants full transparency," seeming to assume that the desire for full transparency is a given for anyone making decisions from models. Transparency serves to enable an understanding of how a model actually works in order to determine if the model should be used for a specific decision. The understanding of a model encompasses, but is not limited to, a model's code, and transparency should include access into practices and decisions involved in creating and validating the model. Moving further along the flow of information decision-making, however, precise information about the models may become less desired, and even unwanted. Comments such as "I trust the people below me" convey the paradigmatic shift that occurs. While details such as model assumptions and uncertainties remain desired, the need for intimate technical knowledge seems to fade. Responses suggest that, even if an actor does not personally require full transparency into a model, transparency should still be available to trusted actors before them in the flow. This suggests a significant point: as actors move further along the flow of information and have less time and ability to personally investigate a model and build their own trust in the model, their trust instead shifts more onto their people to investigate the model for them. In this understanding, the trust for decision-makers is "implicitly on the models, but explicitly on the people."

Understanding of Assumptions and Uncertainty

All models are inherently abstractions of reality that contain assumptions and uncertainties. Models are created for a specific reason and context, and while the assumptions within the model aim to help answer those questions, they also fundamentally create bounds of model applicability. Failure to properly understand the inherent limitations found within a model increases the likelihood of the model being used inappropriately. "All models are wrong, but some are useful," (Box and Draper, 1987) and before any can be useful, their limitations must be understood. As models cannot perfectly encapsulate and relate the situation of interest, uncertainty is fundamentally a part of the results, and uncertainty is also fundamentally a part of determining if the results are appropriately relevant to the decision to be made. This uncertainty must be sought after, understood by the sources of model information, and then passed clearly along the flow of information. There is a fundamental need to understand and express model uncertainties throughout the decision-making flow. Organizational and social dynamics can hinder this expression of uncertainty, however. In some instances, uncertainty about an answer may entail negative stigmas and imply failure to do one's job correctly. Decision-making cultures need to strive to drive out fear of uncertainty expression and transparency. The tragedy of the space shuttle Columbia offers a painful reminder of what can happen when important information is not effectively passed along the decision-making flow (Smith, 2003).

Documentation

Model developers internally carry within themselves the most intimate knowledge of a model's limitations and capabilities. Similar to how modeling is a process of making the internal mental models and expertise found within individuals explicit, documentation is a process of making the assumptions and limitations of a model explicit. Models may very well be validated, even accredited; however, this validation and accreditation are for specific conditions, outside of which the model is no longer valid. Multiple interviews revealed the danger of assuming a model can extend to any context needed when in fact its appropriate contexts of use are much more limited. For a model to have any sort of reuse capability, these assumptions and limitations should be documented in an accessible way so that others can understand how they might appropriately apply the model to their specific situation. Models are built to answer a specific question or set of questions, and the early conceptualizations (e.g. whiteboard drawings) of the model and decisions made in the development process can provide important insight into understanding the model in addition to the documentation of assumptions within the model itself. These conceptualizations, if captured, can provide useful artifacts in the understanding and trust of a model. As models become more complex, documentation of assumptions and capturing of conceptual artifacts and decisions will likely prove crucial in allowing actors to appropriately calibrate their own understandings and mental models of if, and how, a model should be applied to specific decision-making scenarios.

Primary versus Supplementary

Of the experts interviewed, distinctions emerge concerning the primacy of explicit models in the decision-making process. Some view models as clear primary sources in decision-making, others adamantly express that they should only be supplemental sources, and still others present the oft-favored viewpoint of systems engineers – it depends. Those that favor models as a primary source in decision-making point to the benefits of increased knowledge and insight that models can provide if done correctly. Others that advocate for supplementary use emphasize the danger of abdicating the decision-making process to models, and point to the inability of models to capture every relevant factor in a decision. One participant noted an increasing reliance on modeling and simulation (M&S) in decision-making, unfortunately accompanied with the increasing desire to rely on M&S without having to "understand the fundamental processes behind it." The variations in responses serve to validate the non-definitive (yet still insightful) answer of "it depends." Truly, how models are viewed and used is dependent upon the model users and decision-makers, along with the modeling and decision-making context. Well-established and validated physics-based models, for instance, might prove to be a primary source in a decision-making scenario, while descriptive or predictive models that are less conducive to traditional validation may contribute more of a supplemental input within a wide range of other inputs.

Limitations

This study makes abundantly clear that modeling is a sociotechnical issue. Keeping with this understanding, a dichotomy of both technical and social limitations of modeling were presented by those interviewed. The social and non-technical issues, however, proved to be far more prevalent (Table 2). From the technical side, access to appropriate input data proves to be a much cited issue that limits modeling. Without input data that either may not exist or may not be accessible (e.g. for intellectual property reasons), many models cannot realistically be made and validated. Other responses including time and money as limitations make clear that modeling is a resource intensive activity, and does not just happen in a vacuum. While responses indicate that there are very real technical limitations for modeling, such as availability of input data or methodologies needed for accurately modeling certain systems (such as modeling the economy), the vast preponderance of limitations to modeling come from the social, rather than technical, sphere. This provides further evidence that humans are a critical part of the modeling process, and that attempts to improve

modeling must take into account social elements. Many of the limitations stem from a mismatch in understanding of modeling between actors: a disparity of mental models, if you will. While there is no simple answer to fix these issues, forms of education may prove crucial to cultivating properly calibrated understandings of the potentials, limitations, and constraints of models.

Table 2. Limiting Factors to Effective Model-Centric Decision-Making

Technical	Social
• Data • Model complexity • Inadequate methods • Lack of transparency and documentation • Interactivity with models	• Changing preferences of decision-makers • Talent of people • Unwillingness to share models or information • Inertia to change • Communication barriers • Time • Money • Team agreement • Educated leadership • Ability to socialize models within an organization • Skill level • Lack of trust/fear of the unknown • Lack of understanding • Lack of desire to understand • Bad past experiences • Generational differences • Organizational differences

Model Pedigree

Pedigree contributes another important factor in the understanding and trusting of a model, and includes an understanding of the history of where a model comes from. Pedigree is tightly linked with the credibility of the individual or organization that developed a model. Pedigree could include pedigree of the model itself, including information on where its data comes from, how the model has been used, who is using it, and to what affect. An organization that has a history of developing robust and effective models may implicitly carry credibility that lends itself to greater trust. This concern for understanding a model's pedigree highlights the importance of being able to know a model's origin, and who created it. The amount of models generated by various actors and used in various decision-making scenarios is vast, and this generation and use of models produces information that may be useful for informing decision-maker trust in other relevant scenarios. The issue, however, is that there seems to be a lack of consolidation of information concerning what models have been developed and to what results. There may be a need for a form of model curation where specific individuals systematically track model development and use for the purposes of capturing this valuable information that may not be readily available.

Non-Advocate Review

Although models strive to reduce complexity of reality to understandable and workable abstractions, they can still be very complex. Verification ("Did I build the thing right?") and Validation ("Did I build the right thing?") (V&V) are crucial for determining the efficacy and relevancy of a model for decisions (Pace, 2004). Just as skill is needed in model development and use, checks like V&V are required to hopefully catch the inevitable errors. However, effective V&V likewise requires skill and is liable to its own errors. One longtime system architect we interviewed emphasized the importance of utilizing independent experts who can review and render judgments concerning the credibility of results and believability of the data

used. Such a team would be composed of individuals with areas of expertise relevant to the problem. One might view the team as analogous to a forensics team that closely examines the data and code being used and makes judgements that assess the efficacy of decisions made along the flow of model information. Depending on the model and decision-making context, the format and formality of reviews could range from formal, externally-based reviews, to informal, internal peer reviews within a team. Whatever the format, a form of review can serve an important part in the creation of an effective model, and as such, should be a process that is transparent to the decision-makers who are ultimately affected.

Investment Bias and Politics

One individual related the story of a program that involved significant investment in modeling and simulation. When the time came for program decision-makers to make a decision, "they had no choice but to accept" the model's answers "given the resources that were spent." Such a story brings to light the potential bias that investment of time and resources into model development will yield correct and reliable results. Further interviews also revealed a potential for decision-makers to use money as a basis for establishing trust in model results. Money may sometimes offer a useful indicator of model capability; however, no matter how much money is spent on a model, the model is still bounded by the problem space it was designed to solve. Just because large amounts of money were spent on a model does not mean that it is appropriate for the decision at hand. If this issue is not a bias in some cases, then perhaps it may be a political pressure to make a decision based off the model results because of the money spent on model development – if not, the money was wasted. Such a logical fallacy should be countered by a fundamental term of economics: sunk cost. Once money and resources have been spent (sunk) they are gone, and no longer should have any bearing on decisions seeking to promote benefit in the future.

Confirmation Bias

In the words of one respondent: "Quite often, what I see is that decision-makers use models as confirmation bias." This statement reflects one potential pathway for models to be used inappropriately, namely, as a means to further one's own preconceptions or agendas that may be incorrect. Just because a decision-maker's intuition for a solution matches up with a subsequent modeling result does not mean the intuition or the modeling was wrong; in fact, it could be a testament to the decision-maker's experience. However, a senior modeler noted the challenge of guarding against bending a model and results to produce answers desired by decision-makers. Another participant expressed the "amusing thing" that in high-level war game simulations, the war games "almost always" are eventually modified so that your side wins. These interviews reflect the importance for all actors to honestly seek truth while participating in the modeling process. Modeling aims to provide solutions to problems; however, if generated and used to advance one's agenda or to inappropriately confirm preconceived notions, the "solutions" provided may in fact be more damaging than if models were not used in the first place.

Humans Endogenous to the System

Underpinning this study has been the clear and consistent theme expressing that humans are endogenous to the model-centric decision-making process. Many senior decision-makers do not have the bandwidth, training, or time to become technical experts in the models that are used to inform their decisions. How do they trust complicated models? As one senior-level decision-maker put it: "The answer is they trust the people." They trust that the people before them in the model information flow handled the data correctly, created, tested, used and analyzed the model correctly, and expressed the results accurately with appropriate information on uncertainties, assumptions, and limitations. Decision-makers trust that those individuals have the appropriate expertise and capability to understand and address the problem at hand. On the other hand, senior decision-makers also need to have the technical judgment to be able to "sniff out" the wrong

answers, and have a healthy technical competence appropriate to the decisions being made. As systems and their models become more and more complex, the need for skilled and experienced individuals to work within the flow of information seems to be more necessary than ever. Yet the inevitability of aging and retirement guarantees that the experts of today cannot be the experts of tomorrow. Without the right people capable of handling the complexities we are creating, the system will fail, regardless of the technology and innovation we throw at it.

Real-Time Interaction with Models

A final question we asked in the interviews concerned the desirability of being able to directly interact with models in real-time while making decisions. Overwhelming, the respondents view interactivity with models as highly desirable. After all, many decisions involve asking "what-if" questions about the model, and direct interaction could serve to gain insight, build intuition, and speed innovation without needing to go through other human actors. This support for model interactivity also comes tempered with caution from some individuals, however. Specifically, caution against allowing actors interactive access to models without a calibrated understanding of the model's capabilities and limitations. As related by one individual, in situations without this appropriate understanding, "I can get lots of results real quick, and I can make lots of bad decisions real fast." These interviews make abundantly clear the importance for properly understanding a model and its associated assumptions before determining one's trust and usage of model results. Such an understanding is crucial for effective and appropriate interaction with models. As stated in another interview, "If you make it so fools can use it, fools will use it." So while direct interaction with models may be rightly desired based off its potential benefits, development and deployment of interactive models must also advance in a smart and conscientious manner to ensure that actors are not being set up for failure due to ignorance of their own limitations.

4.5 Discussion

The increase we see in system modeling is driven both by a desire to better understand complex systems and issues as well as by increases in technological and computational capability. Similar to technical modeling in many ways, automation involves increasing automation in systems as advancements in technology allows. Often this increase results in gains of efficiency and safety, yet the history of automation has also shown that humans are not just outside users of systems, but rather are endogenously critical components of the system. Experience has also shown that increasing technological capability for the sake of technical achievement, without proper consideration for the human component, can have dire consequences. Bainbridge (1983) writes about the "ironies of automation," where introducing automation can sometimes increase the workload and complexity of tasks it aimed to reduce. With gains in modeling complexity and capability pointing to a model-centric paradigm of engineering, we should be cognizant of potential "ironies of modeling" where failure to appropriately account for human decision-makers and actors results in worsening decision-making processes we aimed to improve.

4.6 Conclusion

This study aims to generate empirical insight into how human actors interact with and trust models, while also providing a starting point for continued exploration into how human actors and decision-makers trust, perceive, and interact with models. The analysis of the interviews presents considerations for human-model interaction and trust that experts deem important for effective model use and decision-making. These considerations include practices that interviewed experts implement to aid in their decision-making, along with identified challenges and potential mitigations to challenges that can degrade effective model-centric decision-making.

5 AI and Autonomy Considerations in DoD Model-Centric Engineering

Model-centric environments become more complex as they evolve to capture the complexity of the systems they strive to model. Automation, autonomy, and artificial intelligence (AI) will likely play a key role in enabling model-centric engineering to manage this complexity. This chapter does not provide answers for how artificial intelligence and autonomy should be specifically applied within model-centric engineering. Rather, it is designed to facilitate a discussion addressing broader principles of artificial intelligence and autonomy – principles that may also be applied to model-centric engineering. This discussion focuses on foundational arguments, ideas, and frameworks, in contrast with the empirical approach applied throughout much of rest of the thesis. Ultimately, it advocates for using a framework of extended intelligence (EI) to inform decisions affecting the balance of human and automated control within model-centric contexts.

5.1 Introduction

The fields of AI and autonomy have undergone massive strides in recent years, introducing capabilities that will likely prove beneficial for model-centric engineering. Because the terms "AI" and "autonomy" can be misconstrued and misapplied, this chapter seeks to establish a useful framework for thinking about them in a manner than can then be applied to model-centric engineering. One common framework presents artificial intelligence (AI) and intelligence augmentation (IA) as diametrically opposed goals of technological development (Markoff, 2015). This chapter explores the arguments for this framework in an effort to illuminate their flaws, ultimately arriving at an understanding that AI and IA are not as mutually exclusive as they initially appear

Problem solving and decision-making are being increasingly abdicated to artificially intelligent systems. Enabling autonomous capabilities does not *replace* human decision-making, however; rather, it *displaces* it in space and time. Furthermore, one can argue that intelligence does not reside exclusively within individual entities, but is distributed throughout a system of networks. This perspective views individual actors and artifacts of technology as components of a larger extended intelligence (EI). Instead of existing as irreconcilable, dichotomous entities, AI and IA present complementary paths that converge towards a greater and more effective EI.

5.2 AI, Autonomy, and Automation

Any discussion involving AI, autonomy, and automation should first attempt to establish a general understanding of the terms. According to John McCarthy, the man who coined the term "artificial intelligence" in 1955 (Myers, 2016), AI is "the science and engineering of making intelligent machines, especially intelligent computer programs" (McCarthy, 2007). Key to this definition is the meaning of intelligence, which McCarthy defines as the "computational part of the ability to achieve goals in the world" (McCarthy, 2007). Another father of AI, Marvin Minsky, emphasizes the problem solving aspect of AI, which includes solving problems of search, pattern-recognition, learning, planning, and induction (Minsky, 1960). While a full understanding AI's historical development exceeds the scope of this chapter, suffice it to say that AI serves as means of enabling computers and systems to rationally solve complex problems or take appropriate actions to achieve specified goals in real world circumstances. ("Preparing for the Future of Artificial Intelligence"). The scientific field of AI serves as the foundation for enabling the creation of autonomous systems – systems endowed with "autonomy." Definitions of autonomy include "the quality or state of being self-governing" and "self-directing freedom and especially moral independence" ("Autonomy"). Applying this definition of autonomy to man-made systems, an autonomous system would be completely self-governing and free from human control. This literal application of the definition suggests the term "autonomous systems" may be a bit of a misnomer in that no existing "autonomous" systems are

actually free from human control. Portions of a system's capabilities may be free from direct, real-time human control, but no system is purely autonomous. The idea of unbounded autonomy is nicely captured by a member of the Google car project, Brad Templeton, who asserts that "[a] robot will be truly autonomous when you instruct it to go to work and it decides to go to the beach instead" (Markoff, 2015). As argued by MIT professor David Mindell, "the machine that operates entirely independently of human direction" – in other words, a truly autonomous system – "is a useless machine" (Mindell, 2015). Rather than adopting such a literal understanding, a Defense Science Board report on autonomy defines it as "*a capability* (or a set of capabilities) that enables a particular action of a system to be automatic or, within programmed boundaries, 'self-governing'" ("The Role of Autonomy in DoD Systems"). For the purposes of this thesis, autonomy will similarly be understood as a capability, with AI serving as an enabler of autonomy. Automation may be an outcome of autonomy. While it has varied meanings in different contexts, automation essentially "occurs when a machine does work that might previously have been done by a person." This can range from simple mechanical automation, like automatic transmissions in cars, to complex tasks requiring autonomous capabilities, such as automating the entire driving process. While various definitions of varying complexity and specificity exist for these terms, this chapter will understand AI as an enabler of the capability of autonomy, which in turn enables automation of tasks.

5.3 Artificial Intelligence (AI) versus Intelligence Augmentation (IA)

In *Machines of Loving Grace: The Quest for Common Ground Between Humans and Robots*, John Markoff follows the development of two contemporary, yet markedly different pursuits of technology: artificial intelligence (AI) and intelligence augmentation (IA). As standard-bearers for these categories of research, Markoff describes John McCarthy, who sought to replicate and replace humans through AI, and Doug Engelbart, who championed, "[a]s much as possible, to boost mankind's collective capability for coping with complex, urgent problems" through IA (Markoff, 2015). Markoff asserts that designers insert their values into the systems they create; because they may have significant effects on future societies, these values should be thoughtfully considered. Injected throughout the historical narratives of AI and IA, Markoff warns of the potential dangers that improper development of AI could bring to the world, from economic disaster through technological job displacement, to AI singularity where intelligent robots surpass humans as the superior entities on earth. Ultimately, Markoff argues for pursuing IA development over AI. Although he offers a compelling narrative, the dichotomous relationship between AI and IA is overly simplistic.

The dichotomy between AI and IA may serve a useful means for capturing differences in opinions between researchers in either fundamental camp, but its oversimplification fails to appreciate the potential benefits of an operationalized approach that combines the two. A more nuanced understanding of the premises that motivate each side suggests that an answer may be found in a gray middle. The boiled down goal of AI is to "replace the human." In reality, however, AI does not so much replace the human as change the role of the human. According to Mindell (2015), "[a]utomation changes the type of human involvement required and transforms but does not eliminate it." This statement is illustrated by the example of the glass cockpit and associated automation, where the pilot's role was changed, but not eliminated, by automation (Endsley, 1996). Changing the pilot's role allowed for a reduction in manpower, as seen in the removal of navigators from airline cockpits, but by no means did it lead to the removal of human will exerted upon the aircraft. Rather, automation changed how humans project that will in the accomplishment of flying. Even during an automated landing without pilot input, a commercial airliner is "landing under the control of the programmers who gave it instructions months or years earlier" (Mindell, 2015).

All machines and pieces of technology are functions of human design and intent; what varies is where you abstract that intent through space and time. Imagine, for example, an autonomous flying window washer. Mr. Smith typically washes the windows of a skyscraper, yet along comes an autonomous flying machine that can accomplish all of Mr. Smith's window washing tasks at cheaper costs and greater speed. The AI machine technically replaces Mr. Smith. If you zoom out a bit more, however, we now see Mr. Smith in charge of multiple "autonomous" window washing machines that allow him to wash the windows of multiple skyscrapers at the same time. From this perspective, the machines have replaced Mr. Smith in one sense, but have also augmented his ability to wash windows in an efficient manner. The word "autonomous" is in quotes because, while able to carry out certain tasks without direct human control, all "autonomous" machines are functions of human decision-making and intent. Mr. Smith must tell the machine which windows to wash, or it will not wash windows. Autonomy allows us more options for where we direct human intention in space and time. The wrapper of human intention surrounding AI systems is what renders technological systems useful.

In contrast to AI, IA seeks to "augment" a person's ability to accomplish tasks. This human-centric approach may be closer to an appropriate role of technology than the extreme opposite view of AI, but is also incomplete. The desire to augment human ability should not merely aim to augment the abilities of individual actors, but rather to augment human will and intent for the overall system. Replacing individual actors with automation may sometimes be a more effective way to augment human will and intent overall. An overly narrow mindset seeking solely to augment individual actors in their specific tasks could foreseeably fail to optimize achieving system goals. This could be the case if keeping individuals in current (even technologically augmented) roles creates limitations that would otherwise be removed if the individual humans were replaced by AI. Again, such a removal of individual actors by no means removes the human will and intent from the system; it rather changes how human will and intent manifest themselves. Accepting these explanations of AI and IA leads to an understanding that the supposed "irreconcilable differences" between AI and IA create a false dichotomy.

5.4 Mental Model Calibration

Colonel John Boyd's conceptualization of the decision cycle as an observation-orientation-decision-action (OODA) loop proves useful when considering the interactions between autonomous systems and humans (Boyd, 2005). An entity used by humans to "act" for them is the definition of a tool. In order to accomplish goals, humans project will and intent upon tools – be they biological, mechanical, or digital – to achieve those goals. It is the "decide" piece that becomes the real issue with autonomous systems, as it may appear that humans are shifting the decision-making responsibility to robotic agents. Automation does not remove human will from a system, however, but rather displaces it in space and time. Where human will may have previously manifested itself through a physically present actor, automation displaces the human will governing a system's "decision-making" to the programmers who developed the automation. In reality, an autonomous system does not make free-willed decisions on its own accord, but rather works within the bounds of human decisions encoded within it in the past. In support of this idea, Mindell (2015) argues that "we must deeply grasp how human intentions, plans, and assumptions are always built into machines." Interactions with automated machines and systems are interactions with "designers and programmers who are still present inside it – perhaps through design and coding done many years before" (Mindell, 2015). A challenge lies in coordinating the goals of a system operator and those who created the system. When operators appropriately understand the goals, assumptions, and limitations found within a system's automation, they can project their will upon the system by using it within its appropriate bounds. Mental models of both the programmers and the operators must be considered when operating an automated system.

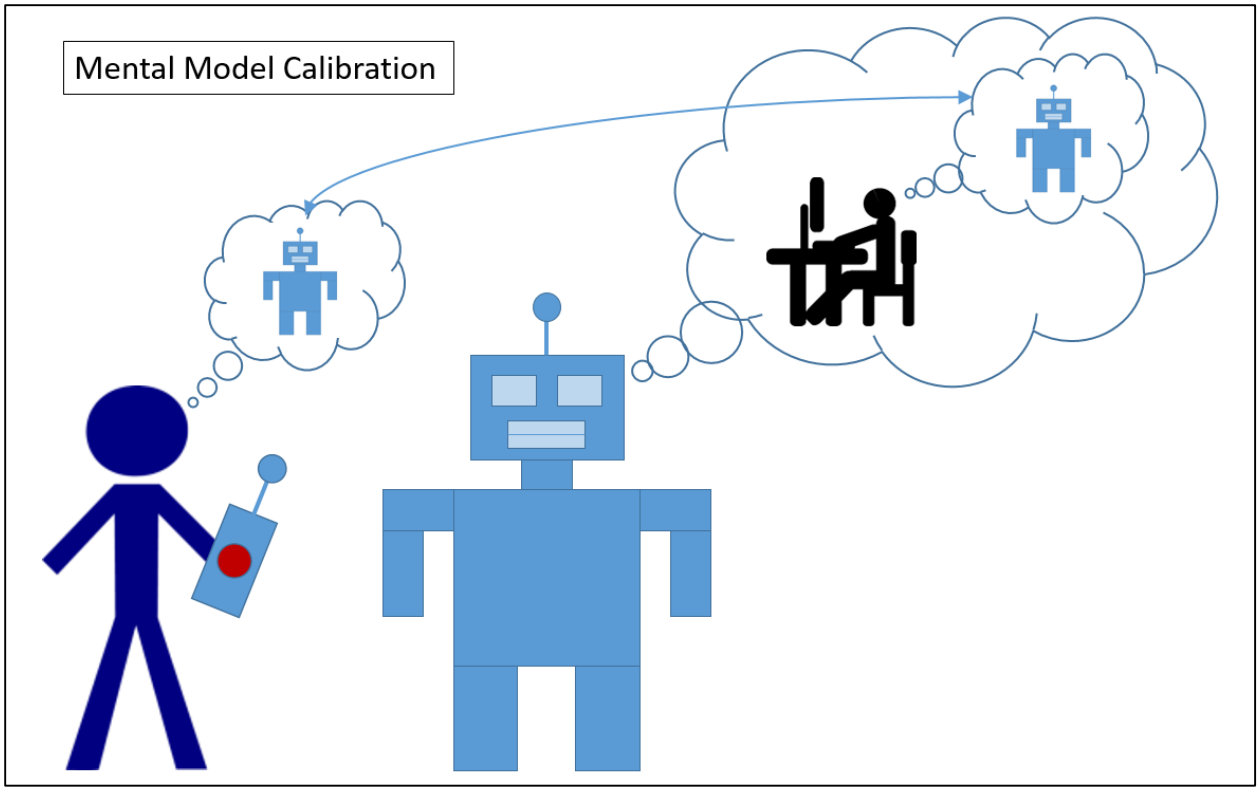


Figure 9. Mental Model Calibration: Automation

Figure 9 presents a schematic to conceptualize the idea that multiple mental models of human will and intent are present during operation of an "autonomous" system. A robot's understanding of its autonomy is fundamentally a product of how the developers programmed the robot to act in various circumstances. When the robot acts, albeit without any direct control, it is fundamentally acting out the directions coded into it by the programmers. The robot is still a function of the developers' will and intent, no matter how far removed in space and time. This intent encoded within the robot is effectively the programmers' mental model for how the robot should react in various circumstances – in effect, coding is the process of making the robot's mental model for how it should act the same as the programmers' mental model. In the case of machine learning, where computers learn and evolve without explicit programming, this projection of will and intent does necessarily specify exactly how to do things, rather, developers specify the goals the AI seeks to accomplish. So while the AI may learn and evolve on its own, this learning still takes place within the bounds of the overall human goals coded within it. The programmers' intent is only one of the human intents involved, however; the other is the intent of the operator who chooses to deploy the robot to accomplish some task (assuming the programmer and operator are not the same person). By choosing to deploy the robot, the operator exerts human will on the system, will that is shaped by the operator's mental model of how the robot will operate. If the operator's mental model is not the same as the programmers' mental model, however, the robot will not perform as desired. The mental models must be calibrated for the intents to be aligned. There exists a dual responsibility, therefore, on the programmers and the operator. The programmers must make the robot's operation clearly understandable and accessible so that the operator's mental model can be calibrated. The operator likewise must appropriately calibrate his or her own mental model before deploying the robot. While "autonomous," the robot's actions are still very much

a product of human will exerted by both the programmers and the user. The user is not so much abdicating real-time decision-making responsibility to the robot as to the programmers who developed the robot.

5.5 The DoD's View of AI and Autonomy

Currently, AI and autonomy represent foundational building blocks for the furtherance of the DoD's Third Offset Strategy. Initiated in 2014 with the goal of "preserving the peace, not fighting wars," the Third Offset Strategy centers around "combinations of technology, operational concepts, and organizational constructs" (Work, 2016). This strategy aims to promote peace through strong conventional deterrence by maintaining military and technological superiority over potential adversaries like Russia or China (Work, 2016). Intended to counteract, or "offset," the conventional numerical superiority of the Soviet Union during the Cold War, President Eisenhower's First Offset Strategy of the 1950's focused on the buildup of nuclear weapons. Secretary of Defense Harold Brown created a Second Offset Strategy of the 1970's that focused on "precision-guided munitions, stealth, and intelligence, surveillance, and reconnaissance (ISR) systems" (Walton, 2016). According to present-day Deputy Secretary of Defense and bureaucratic leader of the Third Offset Strategy initiative Robert Work, the "technological sauce" of the Third Offset Strategy will be "advances in Artificial Intelligence (AI) and autonomy" (Work, 2016). While the technology of AI and autonomy will be crucial to the success of this strategy, technology is not an end in and of itself, but rather exists to "empower humans" and to "[make] the human operate better" (Work, 2016). While AI and autonomy are a critical part of this "human-machine symbiosis," humans "will always be the ones who make decisions on lethal force" (Work, 2016). These comments by Work indicate a high-level understanding that autonomous systems are functions of human direction and intent, yet also promise a commitment to maintain real-time human decision-making when human lives are on the line.

Work's speeches present a top-level vision for AI and autonomy, while reports from the Defense Science Board (DSB) offer greater insight and detail into how AI and autonomy are viewed within the DoD. The 2012 DSB report on autonomy strongly disavows the idea of AI completely replacing humans with statements such as, "all systems are supervised by humans to some degree, and the best capabilities result from the coordination and collaboration of humans and machines" ("The Role of Autonomy in DoD Systems"). Instead, human-machine collaboration provides value by "extend[ing] and complement[ing] human capability" ("The Role of Autonomy in DoD Systems"). Such an approach indicates an understanding that AI should not be intended to displace human authority and that human will and intent are inescapably embedded within autonomous systems: "there exist no fully autonomous systems, just as there are no fully autonomous soldiers, sailors, airmen or Marines" ("The Role of Autonomy in DoD Systems"). That being said, the 2012 DSB report does not view AI and IA as a conflictual dilemma, but rather as a collaborative opportunity, seen by the understanding that both AI and autonomy play a role in augmenting human ability. Indeed, a 2016 DSB report concludes that "autonomy—fueled by advances in artificial intelligence—has attained a 'tipping point' in value," and that the DoD "must take immediate action to accelerate its exploitation of autonomy" ("Summer Study on Autonomy").

5.6 Extended Intelligence (EI)

Limiting the discussion to AI, IA, or even a combination of the two propagates an assumption that intelligence resides within individual intelligent actors, both natural and artificial. Intelligence does not begin and end within boundaries of skin or code, however. Hutchins argues that cognition (a term essentially synonymous with this chapter's use of "intelligence") is "distributed across a social network," rather than being exclusively residing within individuals (Hutchins, 1995a). This theory of distributed cognition views individuals as one of many various cognitive components that form the "information properties of a larger system" (Hutchins, 1995b) Essentially, cognition and intelligence are functions of interactions between

individuals, machines, and artifacts within a system, each carrying meaning and contributing to the distributed intelligence overall. Intelligence, therefore, is not so much a static, individual state, but rather a dynamic "cultural process" that "takes place both inside and outside the minds of people" (Hutchins, 1995a). Constraining the understanding of cognition to individual beings falsely ascribes to them properties that result from interactions between individuals and tools – properties they do not possess on their own (Hutchins, 1995a). Such a constraint fails to examine the processes of cognition and intelligence as they actually exist, and, as such, fails to understand how one might best improve upon the system's cognition. The DoD would do well to more explicitly acknowledge and explore the realm of collective intelligence within its various organizations.

Returning to AI and IA, how would they fit within the theory of distributed cognition? Joichi Ito of the MIT Media lab agrees that "intelligence reside[s] beyond any single mind," resulting in a "networked intelligence that transcends and merges humans and machines" (Ito, 2016). In such a network of intelligence, machines extend human minds into a greater intelligence, and AI serves to augment this networked intelligence. Within this networked, collective, distributed intelligence, all actors and tools contribute to the intelligence as a whole. In Ito's framework of extended intelligence (EI), AI is just another actor further extending the intelligence of the system (Ito, 2016).

5.7 The Role of Autonomy in DoD Model-Centric Engineering

As a part of the transformative shift to model-centric engineering, Blackburn et al. (2017) envisions automation playing a key role in the modeling process. Within this vision, automation "has the potential to subsume the entire" workflow process in a manner that is both "autonomous and adaptive" (Blackburn et al., 2017). As systems and their models become increasingly complex, AI and autonomy present potential capabilities for managing this complexity. How to think about introducing AI into the modeling process, therefore, presents a relevant topic of concern.

As discussed earlier, multiple mental models from different actors exist within the operation of automation. The code within an autonomous system fundamentally represents the programmer's mental model of how it should act or the goals it should pursue in various circumstances. An operator's understanding of automation in relation to a specific situation represents a mental model. Similarly, within a model-informed process, there are many different models, both explicit and mental. First, there is an explicit model that provides information to be used in a decision-making context. Next, there is a decision-maker's mental model of how the model and its information may or may not be valid within the problem's context. Automation would add an additional layer of complexity to this process by introducing the model that guides the automation. The modeling system would be comprised of the explicit model, the automation's model, and the decision-maker's mental models of each. Figure 10 illustrates the idea that multiple models exist within a model-informed decision-making process. The explicit model is a representation of the developer's mental model of what the model should do. All other actors within the decision-making process have their own mental models of the explicit model. In order for the model to be appropriately understood and used, all of the mental models must be calibrated to the developer's mental model for the explicit model. Calibration can occur individually with each actor directly calibrating to the original mental model, but because individuals often build understanding through interactions with others, calibration may also occur between individuals.

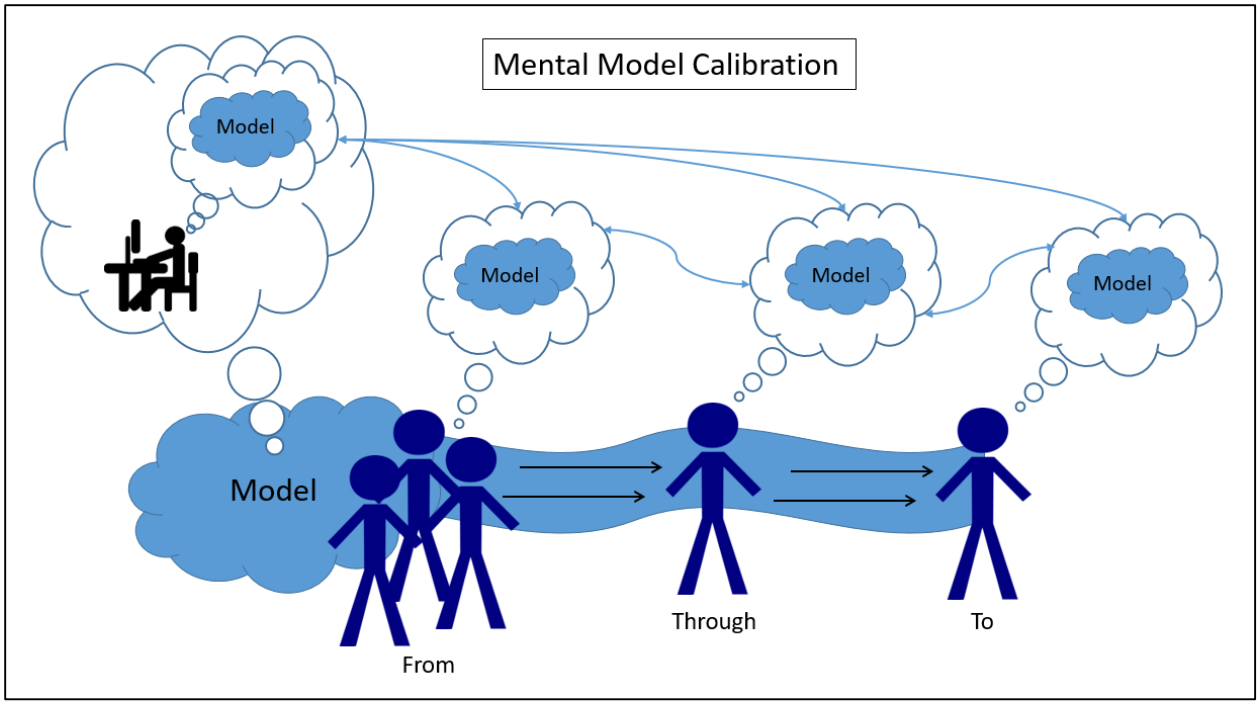


Figure 10. Mental Model Calibration: Model-Informed Decision-Making

Competent Experts

Models do not participate in the *act* piece of the OODA loop, but they are very much a part of the *observe* and *orient* pieces. AI presents the opportunity for models to be granted portions of the *decision*-making capability as well. With this in mind, designers and decision-makers must understand how much of the decision-making piece is granted to models, and ultimately to those who created the models. As mentioned earlier, there exists a dual responsibility between model users and model developers to calibrate mental models – model developers must make the model understandable and accessible, and users must make sure they attain a properly calibrated understanding of the model. This does not necessarily require making models accessible for everyone, however. The 2012 DSB report on autonomy likens autonomous systems to "soldiers, sailors, airmen or Marines," who have autonomous capabilities while still functioning within a broader wrapper of human will and control. Taking this analogy further, the development of effective "autonomous" human warfighters often takes years of training and resources. Similarly, the commanders in charge of those warfighters only achieve that responsibility after years of experience. In regards to model-centric systems, those granted the responsibility to develop or use those systems should only be individuals who are adequately prepared to assume such responsibility. Models should be carefully and intentionally designed to meet the needs of the moment, but the decision-makers and the system itself must also to be trained and shaped in such a manner as to ensure that the models are used appropriately. If a decision-maker does not have the technical expertise to appropriately calibrate his or her mental model with that of those who designed the models, that decision-maker should receive specific training needed to attain that expertise, or should utilize the assistance of other experts before making decisions.

Shanteau (1992) explores what it means to be competent experts, "those who have been recognized within their profession as having the necessary skills and abilities to perform at the highest level." A key area of research is understanding "[w]hat factors lead experts to do well and what factors lead them to do poorly" (Shanteau, 1992). Various factors contribute to whether an individual can be deemed an "expert," including

domain knowledge, psychological traits, cognitive skills, and decision strategies (Shanteau, 1992). These all point to relevant internalized factors within individuals that affect "expertise." However, Shanteau (1992) also points to another "crucial, but often overlooked" factor: task characteristics. That is to say, experts are only competent within a specific range of tasks, and therefore, "the competence observed in an expert depends on the task" (Shanteau, 1992). Much like automation and models, the competency of human experts is both bounded and highly dependent on the task at hand. With this line of thinking, experts should not be "described generically," as one expert may perform competently in one set of circumstances and incompetently in others (Shanteau, 1992) This advances an interesting point: just as decision-makers should not assume that a model or autonomous system is appropriate for a decision or task at hand, one should not assume that a decision-maker is appropriately prepared to handle the task and decision assigned.

Models and Extended Intelligence

Intelligence is not limited to individual actors; rather, intelligence should be viewed as the collaborative intelligence of the system as a whole where tools and technologies serve to extend overall intelligence. The intelligence surrounding a model is highly distributed within the decision-making flow of information. Modelers know the data and assumptions used to create a model in question. Analysts know how to manipulate and work a model to achieve information required from superiors. Senior decision-makers have the high-level perspective of the challenge that must be solved. Program managers have information from both the senior decision-maker side and the analyst side that can be used to communicate between those various levels. Much of the intelligence of the system is built into the individual actors within the decision-making process. However, intelligence is found not just within individuals and the model itself, but also within the artifacts of its creation. Documentation of assumptions and limitations of a model and information about who developed it, the source of the data that helped create it, and the success of the model in the past are factors that are integral to calibrating an appropriate understanding of how the overall system intelligence may be used to solve a problem at hand. Attempting only to augment individual actors in this situation may fail to capture the benefits that computers could add to the decision-making process, yet replacing certain individuals with AI may intractably remove sociological factors that influence the trust and ultimate use or disuse of a model in decision-making. Consideration of how to incorporate AI within the modeling process should take care to extend, rather than detract from, the overall intelligence within the decision-making process.

5.8 Conclusion

The general principles concerning human interaction with physical robotic systems also applies to a discussion surrounding autonomy within modeling. Of primary import is the understanding that autonomy is always a function of human will and intent. With this understanding in mind, system designers can trace back to where intent is exerted and determine if that is where they want responsibility to reside. Once responsibility has been appropriately delegated, effort can be focused on how to achieve effective use of autonomy. Key to effective use is the process of calibrating the mental models of the various actors involved. Calibration carries a dual responsibility: those designing the system must create it in such a way that it is understandable, and those deploying the system have the responsibility to make sure that they are appropriately qualified and that they have appropriately calibrated their understanding of the system to the intent of the designers.

Hutchins (1995a) writes that "one of the primary jobs of a theory is to help us look in the right places for answer to questions." A theory positioning AI and IA at fundamental odds with one another fails in two areas. One, in that AI and IA are not mutually exclusive. Full autonomy is a myth: AI and the autonomy it enables cannot escape the "wrapper of human control" of human will and intent that is projected upon any

system, regardless of whether it is “autonomous” (Mindell, 2015). IA, which seeks to augment human intelligence, can actually use AI to physically replace humans in the augmentation effort. The DoD exhibits a grasp of these concepts in its pursuit of human-centric AI and autonomous system development and would do well to apply them as it pursues the possibilities of AI in its pursuit of model-centric engineering. Such an understanding promises far greater success than any approach limited to just AI or IA. A second challenge is that the focus on AI vs. IA carries an underlying assumption that the intelligence of system is found within individual entities. Rather than primarily considering AI or IA, this chapter argues that the focus should center on EI. The theory of distributed cognition and EI aims to direct attention to the influential cognitive properties that take place through the interactions between the components of an intelligent system. The intelligence found within systems is larger than the sum of the intelligence found within individual components of the system. It is the networking of the collective, distributed intelligence of the various system components, human and artificial alike, that produces the system intelligence. Any approach attempting to extend a system’s intelligence of any sort would be wise to take to take EI into account.

6 Guiding Heuristics

This chapter distills the research from the previous chapters into twenty-nine descriptive and prescriptive heuristics for enabling effective human-model interaction and model-centric decision-making. These heuristics emerged from the voice of the experts interviewed, as well as case studies and literature analyzed.

Heuristics can be thought of as prescriptive theories that offer "suggestions to facilitate more extensive search for useful possibilities and evidence" (Baron, 1988). Maier and Rechtin (2000) defines heuristics as "sophisticated abstractions of lessons learned from experience." Such condensed and codified abstractions of practical experience should be thought of as tools to assist individuals in design and decision-making. There may be hundreds of potential heuristics available to a decision-maker, "but only a few are needed at any one time and for a specific job at hand" (Maier and Rechtin, 2000). Heuristics are "guides along the way," but also "must be used with judgment" (Maier and Rechtin, 2000). There is an art to this discernment, not in evaluating the wisdom of heuristics themselves, but "in the wisdom of knowing which heuristics apply" to a given problem (Maier and Rechtin, 2000). As dual-process theory makes abundantly clear, heuristics, when improperly applied, can lead to systematic biases. As with any tool, discernment concerning applicability of these generalizations to specific contexts is required.

6.1 Heuristics for Human-Model Interaction and Model-Centric Decision-Making

Heuristics are grouped under six categories: (1) Designing Models for Human Use, (2) Using Models in Decision-Making, (3) Sociotechnical Considerations, (4) Context and Assumptions, (5) Transparency and Trust, and (6) Mitigating Biases. Numbering of the heuristics is for identification purpose only, and is not intended to imply order of importance.

Designing Models for Human Use

1. **Humans should not be forced to adapt to models, rather, models should be designed for humans.**

Technological development will enable models to increase in complexity and capability; this capability may not equal an increase in effectiveness, however, if humans are not appropriately considered. Models represent a means for extending the intelligence of an organization as a whole and should not detract from human capabilities and intelligence. Humans have cognitive and perceptual limitations that limit the amount and type of information they can effectively comprehend and use to make decisions. Designing for humans requires understanding their capabilities and limitations within the context of the system, so that the model intelligence can extend the overall system intelligence.

2. **Model design must consider users and how the design will influence user behavior and decisions.**

Humans are adaptive creatures that can learn new skills, but we also share predictable limitations and behaviors that can be shaped by the environments in which we operate. Model-centric designers should not expect users to adhere to normative, or ideal, models of behavior and decision-making. Rather, developers carry the responsibility to seek out and understand how users will react and behave within certain contexts, and shape the design to extend human strengths while mitigating weaknesses. Instead of placing the responsibility for effective decision-making squarely upon the shoulders of the final decision-makers, this view distributes responsibility to designers, who should have a sense of accountability knowing that their decisions will affect user behavior.

3. **Applying user-centered design will enable model-centric environments to give users what they actually need, not just what they want.**

Model-centric environments should be created in a manner to efficiently and effectively extend the intelligence of individuals and organizations. To accomplish this, the design and structure of these environments should focus on how best to meet the needs of users by extending strengths while minimizing weaknesses in cognitive and perceptual ability. This paradigm of design includes both technical and organizational considerations, and requires an intimate understanding of the individual users and their capabilities, limitations, needs, and preferences. Individuals may be subject to preference-performance dissociation, however, in which they want something that actually leads to poorer performance. For example, a decision-maker may want to use certain capabilities and models, but if the individual is not properly trained and equipped to interact with them, the organizational practices should prevent their use without a properly calibrated understanding and trust.

4. **Tailorability allows users to adapt the environment to meet the needs of a decision at hand.**

An initial design cannot predict all of the situations that decision-makers and users will face. Tailorability allows individuals to modify the environment in a manner to meet the needs of a current situation. This tailorability should be viewed in light of the heuristic on user-centered design, however. A good design includes a delicate balance between tailorability that allows modification, and designs that adhere to predetermined normative ideals.

Using Models in Decision-Making

5. **Models do not have agency -- the ultimate responsibility for decisions must be upon humans.**

The ultimate decision-making authorities within model-centric environments are people, and blame cannot be placed upon models for poor decisions. With this in mind, model developers, users, and decision-makers have the responsibility to ensure that models are properly understood and appropriately used. Increasingly complex and automated modeling environments may make it easier to attribute agency and responsibility to models. Individuals should be aware of the potential for improperly diffusing responsibilities for decisions upon models, and policies should clearly establish the responsibilities for which individuals are held accountable. As long as people are the final decision-making authorities, policies should encourage and require strict and clearly defined responsibilities and accountability. This structural accountability helps normalize a culture that promotes systematic and analytical thinking to avoid bias-prone, heuristic-based decision-making.

6. **Models should be treated as tools, not as agents. They are all products of human design and intent.**

Models are ultimately tools that individuals and organizations use to contribute to decision-making, and do not make any decisions themselves. All models are products of human will and intent, and are used as instruments of users' intent. The inclusion of artificial intelligence and autonomous capabilities within the modeling process does not change this fact. Models incorporating automation are designed and created by people, and other people choose how to employ and maximize utility from these products. Automation shifts the direct control of certain processes in space and time from individuals who previously directly interacted with those processes, to developers who created the automation. This places a responsibility upon developers to design automation in a safe and understandable manner, while those who implement the autonomous capabilities have a responsibility to understand the automation and apply it in a proper manner. Automation does not exist in a vacuum; it transfers responsibilities for system actions and correspondingly changes the behaviors of individuals within that system. Inclusion of automation, as with any system

change, should be preceded by detailed analysis of possible effects, and vigilant monitoring and adaptation following implementation.

7. **Not all models and modeling contexts are created equal – some models may be viewed as primary sources in decision-making, while others may be viewed as supplementary.**

Well-defined problem contexts may allow for accurate modeling with low uncertainty, yet the same model applied in a different context may have increased uncertainty. Certain types of models may carry greater weight in decisions than in others. For example, a physics-based model validated over years of use may take a more primary role in decisions than a descriptive model based on subject matter expert (SME) input. Regardless how much weight is given to a model, humans are still the ultimate decision-makers carrying the responsibility for decisions, including the decisions concerning what models to use and how. The decision to use models should include an appropriate understanding of how much weight model outputs should be given.

8. **Be wary of modeling for its own sake -- poorly applied models can make decision-making outcomes worse.**

Model-centric engineering aims to enable more effective and efficient decisions. This requires that both models and users be capable and appropriately qualified for the decision at hand. Models may be highly capable, but if they are not understandable they may be used inappropriately, resulting in poor decisions and outcomes. Additionally, models should not be created for the sake of having models; rather, they should have a purpose and capabilities relevant to the decision at hand. Regardless of the model being used, if decision-makers are not properly equipped to handle information from a specific model, they may inappropriately apply the model to situations out of its inherent bounds and limitations.

Sociotechnical Considerations

9. **Model-centric engineering is fundamentally a sociotechnical process composed of human actors interacting with models, model-generated information, and one another.**

How to model a problem or opportunity in a useful manner is a sociotechnical issue that requires understanding not only the technical side of how to model, but also the social side of how human actors perceive, trust, and use models. A highly technical and capable model may offer value if used appropriately to solve a certain problem, but it will not actually be used if the individuals involved with the problem do not understand the model and its potential value. If the goal of modeling is for the model to actually provide value, then a holistic perspective considering both the technical and social components of the problem and the individuals involved is necessary.

10. **When decision-makers can directly engage model developers as needed, they will be more likely to use a model's results.**

Model development and subsequent analysis can be a complex process that provides unintuitive answers for those not intimately acquainted with the model. If a model's results are not understood or trusted by a decision-maker, they will likely not be used, regardless of whether or not they could have actually been useful. Communicatory pathways that allow for iterative, two-way communication between decision-makers and upstream actors helps ensure that the model is being developed to answer the right problem while also providing decision-makers with a more sophisticated understanding of its capabilities and limitations. Iterative communication involves decision-makers providing guidance to model developers, who in turn update decision-makers on preliminary design and results and receive feedback as needed. Rather than giving the problem to model developers and trusting they understand the problem and will solve it correctly, iterative communication allows decision-makers to solve the problem by employing

model developers as extended agents of themselves. Individuals with technical understandings of a model can inadvertently speak over the heads of decision-makers who lack the same understanding, and vice versa, essentially creating language barriers. These barriers require a "translation" so each side can understand one another. Therefore, a dual responsibility exists on both the parts of decision-makers and upstream actors to ensure that information about models and the problem at hand is communicated iteratively and effectively.

11. **The value of a model-informed decision-making process stems from the flow of model-generated information *from*, *through*, and *to* human actors involved in making and supporting decisions.**

A model does not provide value by existing, but rather by producing information that can be used by individuals and decision-makers to better understand a specific problem. Often, information flows *from* initial actors that directly interface with a model before flowing *through* other individuals within the decision-making process and ultimately *to* final decision-makers. All of the actors within this model-generated flow of information must have a well-calibrated understanding of the model if the information is to be effectively understood. How, and through whom, information flows should be understood so that an organization can ensure the model-generated information is understood and used in a manner that is appropriate and adds value to the overall process.

12. **Direct, real-time interaction with models is desirable; this should be approached with caution, however, ensuring that decision-makers have the right training and experience necessary to effectively use the models.**

While models should be designed for human users, those individuals must likewise be properly trained to appropriately understand, trust, and use models. Similar to how pilots are not allowed to fly an aircraft without proper training and qualifications, those involved within the modeling process should be properly trained and qualified to use models appropriately. Improper matching of human capabilities to a complex modeling environment leads to suboptimal decision-making. Two interviewees succinctly verbalized this caution. One stated that with such an interactive setup, "I can get lots of results real quick, and I can make lots of bad decisions real fast." The other stated, "If you make it so fools can use it, fools will use it."

13. **Model developers, model users, and organizations have a shared responsibility to combat improper model usage.**

The assumptions within a model fundamentally bound the applicability of the results to a specific problem space. This introduces the possibility for model misuse, disuse, and abuse. A model may be misused if used outside of its bounds to aid in a decision where model-generated information is not valid. Disuse occurs if a model that could provide value to a decision is not used and its value not maximized. Model developers have the responsibility to make capabilities and limitations of models salient and understandable, while model users have the responsibility to ensure they know this information before using or not using a model. Organizations have the responsibility to prevent model abuse – the improper use of models in general. While models are useful tools, they are not a panacea, and organizations should not treat them as such. Models should be created as means to extend the intelligence of the organization as a whole, and should not be abused.

<u>Context and Assumptions</u>

14. **Assumptions may be made within a model, but should not be made about a model.**

Assumptions are a fundamental part of modeling, but decision-makers and users should not make assumptions about models under consideration. The assumptions within a model bound its applicability,

and decision-makers have the responsibility to ensure the model is appropriately applicable to the problem at hand. Models need to make assumptions. Decision-makers do not need to make assumptions about the model.

15. **Before any model can be useful, its capabilities and limitations must be revealed and understood.**

All models are inherently abstractions of reality that contain assumptions about the modeled system, and these assumptions limit the applicability of where and how the model can be used. Numerous empirical examples show that poorly understood and applied models have led to programmatic challenges and failures. An improper understanding of a model's limitations may lead to decision-makers inappropriately applying it beyond its limitations. Conversely, if a model's capabilities are not fully understood, decision-makers may not use the model to its fullest potential.

16. **Never assume a model that is applicable in one context will be applicable in another context. Models are created for specific reasons and contexts, and those assumptions fundamentally bound a model's applicability.**

Although all models contain assumptions, these assumptions must be valid for whatever problem context is being considered. A model may be insightful and valuable within one problem context, but the assumptions built into the model may not be valid within some other context. The history of a model's performance may offer a useful indicator for a model's capabilities and value; however, this performance record should be accompanied with an understanding of how, why, and where the model was applied. A model may offer strong explanatory and predictive power in certain contexts, but not in others. Evaluating a model's applicability should not just consider if it has been validated and verified, but in what contexts it has been validated. Using a model outside of its inherent bounds may lead to model results that are inappropriate for the problem under consideration.

17. **Model documentation should make the assumptions and limitations of a model explicit. Without this information, a model is not usable by anyone other than the model originator.**

Model developers carry the most intimate understanding of a model's assumptions and limitations. If decision-makers are those other than the modelers, however, assumptions and limitations must be clearly documented so that others might calibrate an appropriate personal understanding of the model. Documentation should not only capture the assumptions built within the model, but the assumptions made about the model itself. Conceptual "white board" artifacts created early in a model's development can offer insight into decisions made about a model, including what problem contexts the model is designed for. With this line of thinking, both decisions made regarding assumptions within the model's code and the decisions made before the model was even designed should be documented. If these assumptions and limitations are not documented and accessible, users and decision-makers will not be able to appropriately calibrate their understanding and trust of the model, which makes it unusable.

18. **Modeling is not only about creating the models, but also ensuring their appropriate use. Appropriate use included matching a model's capabilities and limitations to a given purpose and context.**

Models serve as a means for updating an individual's mental model, or understanding, of a situation. The addition of further information can create a higher fidelity mental model of a specific situation. If improperly understood, however, the information supplied by models may lead to an inaccurate understanding of a situation and, subsequently, improper decisions. Key to understanding a model's information appropriately is not only understanding the model's capabilities, but also the limiting

assumptions found within it. Therefore, models should not be viewed as an end in and of themselves, but rather as a means for bolstering an individual's ability to make good decisions by providing a means for developing a more well-informed understanding of the situation.

19. Large amounts of time and money invested in a model do not necessarily mean the model is appropriate for a given situation.

The time and money invested in a model's development may offer a useful indicator of a model's quality, but does not guarantee quality – it is possible to make a very large and expensive model that does not work well. Even if a model is high quality, this does not equal applicability. A highly capable model may prove useful in certain contexts, but has fundamental limitations that make it inapplicable in other contexts and problems. Rather than relying on the heuristic of resources invested in a model to determine its efficacy for a problem at hand, decision-makers should determine if it is actually applicable.

Transparency and Trust

20. Calibrating appropriate trust in models is crucial for enabling effective model-centric decision-making.

Trust is an important factor in determining a model's use; however, this trust may be misplaced. A decision-maker may over-trust a model and use it inappropriately outside of its bounds, or, conversely, a decision-maker can inappropriately distrust a model and not utilize its potential value. Furthermore, as model-centric decision-making is a sociotechnical process, an individual's trust in a model can influence another's trust. Therefore, all actors within a model-centric decision-making process should have an appropriately calibrated trust in the models used, or they may negatively influence one another and decrease the utility of the model.

21. Trust is a sociotechnical construct: you must examine both technological and social factors to understand how individuals develop trust in models.

Individuals within the model-centric decision-making process rely upon various technological and social factors to develop trust in a model. Technological factors include technical information about a model, such as its transparency, uncertainty, and input data. Social factors include the people, organizations, and relationships that shape one's trust of a model. These social aspects could include factors such as the credibility of the people or organization developing the model, reliability of the relationship with individuals recommending a model, or word of mouth within a community concerning a model's performance. Different factors may play greater or lesser roles in developing an individual's model trust process; therefore, these factors should be understood to facilitate appropriate calibration of individual trust.

22. Models should have the capability to be as transparent as possible; however, not every user desires full transparency. Transparency should be tailored to the needs of the specific individual under consideration.

Transparency involves how clearly one might assess a model's functions and understand how and why it operates as it does. This allows individuals to determine if the operation is appropriate for a decision at hand. Full transparency would involve having complete access to a model's code and documentation of the assumptions built within it. While there should always be the opportunity for full transparency, individuals may desire different levels of transparency. For example, high-level decision-makers may only desire transparency concerning high-level model assumptions because they lack the time or training to effectively investigate a model's code. Too much transparency could cause an information overload that obscures the relevant information. Conversely, others may desire the ability to be intimately acquainted with a model's

workings. Transparency, therefore, should always be present, but should be tailored to the needs of the specific individual under consideration.

23. For those without an intimate understanding of a model, trust is often placed more heavily on other individuals than on their personal understanding of the model.

A person's trust in a model can be shaped through numerous technological and social factors. As decision-making responsibility grows, however, individuals may find themselves with less time and technical expertise to build a robust personal trust of a model. For those individuals who do not have the time or ability to develop trust through thoroughly examining the technical details of a model, they rather tend to shift their trust on others to investigate the model for them. This emphasizes the importance of having skilled and trustworthy individuals within the decision-making process. In such cases, by saying they "trust" a model, decision-makers are really trusting individuals who said the model could be trusted.

24. Effective model trust calibration involves a process of determining the applicability and efficacy of the model for a decision at hand.

Calibrating trust in a model involves assessing its capabilities, assumptions, and limitations, and determining if such an abstraction can be usefully applied to a decision-making situation. Proper trust calibration for a model stems from an accurate understanding of its capabilities and limitations. Once those aspects of a model are properly understood, an individual can determine whether to trust it to aid in a specific decision-making context, to use it in another, more applicable, context, or to not use it at all.

25. Appropriate model trust calibration includes the sharing of mental models between actors within a decision-making flow.

A model is the explicit representation of the model developers' mental model of a situation. This mental model contains an implicit understanding of the capabilities and limitations of the model. The other individuals who interact with the model should calibrate their own mental models to that of the model developers, which includes developing an accurate understanding of the model's capabilities and limitations. Improperly calibrated mental models lead to inappropriate decisions on an individual basis, but as modeling is a sociotechnical process, improper mental models may also influence and hinder proper mental model development in other individuals within the flow of information.

Mitigating Biases

26. The path of least cognitive effort may not always be the right path.

Biases stem from using generalizations and heuristics that improve efficiency in decision-making and have predictive power in many situations, but also cause systematic errors in specific situations where the heuristics are not applicable. Heuristics also offer an easier path for decision-making than critical evaluation of a problem. In many cases, following the path of least cognitive effort will increase efficiency, but also increases potential for bias. People are naturally predisposed to be "lazy" and often substitute easy questions for the actual hard ones, even if doing so is irrational. To combat potential biases, individuals within a model-centric environment should be expected to make rational, thought out decisions. Because individuals are naturally prone to follow the easiest routes, designs should consider whether they are promoting or preventing improper cognitive ease. For example, automation bias stems from following the path of least cognitive effort while not critically examining an automated system's actions and suggestions. Any form of automation should be added in a manner that adds beneficial efficiency while still preventing the susceptibility of automation bias.

27. **Increasing the speed of decision-making implies a decrease in time spent analyzing a problem, which in turn increases the chance of biased judgments.**

Humans excel at pattern recognition and efficiently jumping to conclusions in the presence of limited information. These intuitive capabilities prove effective for many day-to-day activities, but they can also lead to systematic biases in certain contexts. Model-centric environments may seek to provide a more intuitive environment that users and decision-makers can directly interact with to build intuition and speed decision-making. However, these environments may make biases more likely by allowing users to quickly match patterns and move forward with incomplete, and potentially biased, understandings of a situation. Complex problems may require focused attention and analysis that take time to fully understand in order to develop an accurate mental model of the situation. While faster decisions are desired if they are effective, the speed itself may set people up for failure by encouraging them to rely upon their fast and intuitive, yet bias-susceptible, judgment, rather than the more cognitively demanding rational and analytical thought processes.

28. **Non-advocate review of models provides a means for raising important questions, challenging assumptions, and detecting biases that model users and decision-makers may fail to see themselves.**

Individuals each have their own unique mental models of situations, so subjecting a model to a diverse group of experts allows it to be stressed in various ways that may reveal weaknesses. The approval of a model's design and results by a diverse group of relevant experts offers a useful indication of model quality and applicability. In general, the more sets of expert eyes you can get on a model, the more likely you are to reveal flaws that were otherwise hidden. Humans are notoriously bad at detecting judgmental biases within themselves, but are better able to detect biased and irrational decisions within others, including the decisions used to develop and use a model. Both internal and external reviews offer formalized pathways for detecting flaws in both model design and use.

29. **Model-centric environments need to be structured to minimize or make evident potential biases and failure modes that users are often unaware of.**

Biases operate in a subconscious manner that makes personal bias detection and mitigation challenging. Strategies to improve individual decision-making through training and education may provide some benefit, but rather than solely focusing on trying to change the human actors, the system should be designed in a manner to mitigate biases through prevention or detection in bias-prone situations. Checklists represent one manner of forcing individuals to adhere to normalized practices and decision-making processes. Non-advocate reviews offer another mechanism for allowing others to catch biases and irrational decisions made by others, which is much easier to do than catching the irrationalities in one's own judgment.

6.2 Heuristic Validation

Maier and Rechtin (2000) suggests the following test for validating heuristics:

> "There is an interesting human test for a good heuristic. An experienced listener, on first hearing one, will know within seconds that it fits that individual's model of the world. Without having said a word to the speaker, the listener almost invariably affirms its validity by an unconscious nod of the head, and then proceeds to recount a personal experience that strengthens it. Such is the power of the human mind."

Following this model of validation, initial validation of the heuristics was accomplished through feedback and critique from two rounds of internal MIT validation. The heuristics were changed and evolved in

response to received critique, arriving at the final heuristics presented in this thesis. Going forward, these heuristics will be further validated by external experts.

These heuristics represent descriptive and prescriptive encapsulations of effective human-model interaction and model-centric decision-making that emerged from this research. All heuristics are limited in their descriptive, prescriptive, and predictive power. They should not be viewed as fully generalizable, universal truths, but rather as tools that system practitioners can apply in appropriate contexts. Following further validation, these have the potential to be fundamental principles for use in educating students and the existing workforce

7 Policy Considerations for Model-Centric Engineering in the DoD Context

This chapter presents policy considerations, specific to the DoD, in its pursuit of effective model-centric engineering. In addition to advocating for continued research into challenges that face human-model interaction and model trust calibration, a general policymaking approach of planned adaption is suggested. Such a framework for policymaking acknowledges uncertainty in the policymaking process, and establishes pathways for gathering new information to reduce uncertainty and create more effective policies. Planned adaptation enlarges the view of the policymaking process into one that explicitly plans for, and works towards, adapting policies as a natural part of a policy's "lifecycle."

7.1 4 C's of Model-Centric Engineering Research

The 4 C's (Figure 11) offer a conceptual model for framing the structure and purpose of this thesis. The "capabilities" represent the primary vision for model-centric engineering, essentially, its desired outcomes. The "competencies" denote the knowledge, skill, abilities, and enabling technologies required to achieve those capabilities. Overall, the most general goal of model-centric engineering is to enable more effective and efficient decisions. A third "C," focuses on developing "cautions" shaped by relevant challenges that could hinder the goals of efficiency and efficacy in system decision-making. This thesis primarily focuses on developing these cautions, culminating with descriptive and prescriptive heuristics for enabling effective human-model interaction and model-centric decision-making. Finally, these cautions feed into "controls" – policy and guidance that shape the capabilities and competencies of model-centric design, development, and use. This chapter offers policy considerations for this fourth "C."

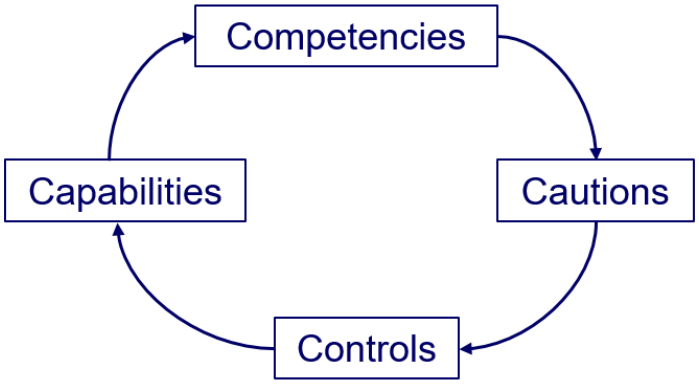


Figure 11. The 4 C's of Model-Centric Engineering Research and Development

7.2 Relevant Policymaker

As established in DoD Instruction 5134.16, the Deputy Assistant Secretary of Defense for Systems Engineering (DASD(SE)) serves as the "focal point for all policy, practice, procedures, and acquisition workforces issues relating to systems engineering, development planning, and related engineering fields within the Department of Defense." Figure 12 illustrates how the Office of the DASD(SE) (ODASD(SE)) aims to develop policies, guidance, and education and training (E&T) for programs based on research and development and interaction with relevant communities. One such research initiative is the Systems Engineering Research Council (SERC), which funds various academic research projects, including the

Interactive Model-Centric Systems Engineering (IMCSE) project, from which this thesis is a product. Research initiatives and projects have been established as a means of promoting research and development to inform policy decisions.

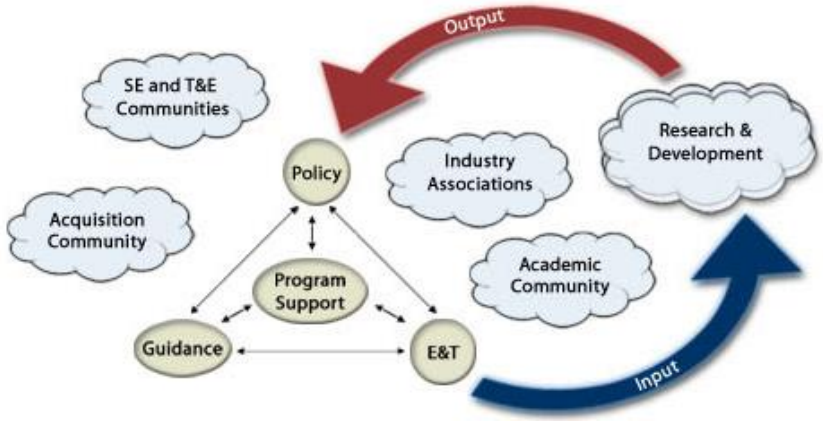


Figure 12. ODASD(SE) Approach to Policy

7.3 Policy Considerations

This research into human-model interaction generates a number of considerations for informing policy and guidance on the development and use of model-centric environments, including its practices and technologies.

Policy Consideration 1: *Humans are critical components of effective model-centric decision-making. Policies should encourage model-centric practices that seek to understand and design for these human actors.*

Understanding the human aspects of human-model interaction is crucial for enabling effective model-centric decision-making. Without appropriately considering and designing for the human element, model-centric engineering will struggle to achieve its desired outcomes. Enabling effective human-model interaction, therefore, is crucial for realizing the value that models and model-centric engineering practice can provide.

Policy Consideration 2: *Policies need to address the unique challenges posed by human-model interaction within model-centric environments, challenges that go beyond those already considered by human systems integration.*

Humans are highly capable, but also have limited cognitive and perceptual abilities. These limitations contribute to errors and biased decisions. While this research makes clear that significant challenges exist when humans interact with abstracted information in decision-making environments, further research is needed to more fully understand the challenges that face human-model interaction within model-centric environments. Human Systems Integration (HSI) is an initiative in the ODASD(SE) which addresses human element considerations throughout the acquisition development process (DDRE). Research into human-model interaction is relevant to the HSI initiative, and vice versa. While human-model interaction has unique aspects itself, HSI knowledge and practices may offer a useful collaboration for model-centric design, development, and use. Chapter 3 provides an in-depth discussion relevant to this consideration.

Policy Consideration 3: *Many factors, both technological and social, influence an individual's trust in models. Rather than merely specifying that individuals should have an appropriate level of trust in models, policies should promote understanding and utilization of these factors to engender appropriate model trust.*

Many different actors reside within model-centric decision-making processes, each having specific sociotechnical factors that shape their understanding and trust of models being used. An individual may over-trust a model and use it inappropriately outside of its bounds, or, conversely, an individual may inappropriately distrust a model and not utilize its potential value. Calibration involves actively assessing the sociotechnical factors relevant to a model in order to develop appropriate trust. These sociotechnical factors need to be understood and leveraged to cultivate appropriate calibration. Chapter 4 provides an in-depth discussion relevant to this consideration.

Policy Consideration 4: *Automation within model-centric environments adds additional complexity to considerations of human-model interaction. Policies should promote active consideration of the effects automation will have upon user behavior before implementing autonomous capabilities.*

Automation changes the role of humans within a system which also has a direct effect upon human behavior. This change can result in both positive and negative outcomes (Chapter 3). Considerations of incorporating AI and autonomous capabilities within model-centric environments should be judged against whether or not they will extend the intelligence of the system. Of key importance is understanding that automation does not replace human will and intent, but rather displaces it to those who developed the logic encoded within the automation. Developers and users have a shared responsibility to ensure automation is implemented appropriately. Automation should also be appropriately transparent so individuals within model-centric environments can calibrate an appropriate understanding of trust in the automation in addition to the overall modeling system. Chapter 5 provides an in-depth discussion relevant to this consideration.

7.4 Planned Adaptation

One primary challenge with policymaking for any new technology is uncertainty. No one knows exactly what the future will hold, and regulators must dig through this uncertainty to the best of their abilities to create policies that maximize social benefits while minimizing risk. In the end, whenever uncertainty surrounds a new technology, policymakers' regulatory decisions represent a "best scientific guess" for an effective solution. This issue presents policymakers, such as governmental officials in charge of overseeing a DoD shift to model-centric engineering, with a dilemma: even though policymakers seek to maximize benefit and minimize risk, any decision they make will undoubtedly fail to accomplish either of those goals fully. For regulatory improvements to occur, new information must be available to inform improvement, and secondly, mechanisms must exist to change existing policies and regulations. McCray, Oye, and Peterson (2009) specifically addresses the issue of risk regulation under uncertain conditions with a term called "planned adaptation". Planned adaptation acknowledges that mistakes will be made when creating regulations under uncertainty, but, rather than fatalistically accepting this fact, also seeks to actively mitigate and improve upon mistakes made. Instead of changing policies in a reactionary manner, organizations committed to planned adaptation have a "prior commitment to subject an existing policy to de novo re-evaluation" in addition to creating a "systematic effort […] to mobilize new factual information for use when the re-evaluation takes place" (McCray et al., 2009). Such mechanisms for planned and informed change, if committed to, create a means for policymakers to continue pushing towards the ever-elusive goals of maximized social benefit and minimized risk.

While planned adaption may be reasonable and readily accepted in theory, planned techniques of regulatory adaptation have struggled to receive substantial implementation within the United States (McCray et al., 2009). Many possible reasons exist for policymakers' reluctance to pursue planned adaptation, including preference for the status quo, fear of losing credibility, and the challenges of enforcing constantly changing regulations. Additionally, money can prove to be a significant detriment to planned adaptation if organizations lack proper funding to consistently review and update regulations. Indeed, situations where new knowledge may not be readily available or useful may rightly deem stability to be more valuable than the costs associated with updating regulations. Planned adaptation may be most effective and appropriate when major uncertainties are present, or when large benefits and costs are at stake – criteria that certainly apply to the realm of model-centric engineering (McCray et al., 2009).

Walker, Rahman, and Cave (2001) provides additional support for planned policy adaptation, noting that policies that deal with complex systems and issues must cope with "profound uncertainties about the future." These uncertainties all but guarantee the "pragmatic reality" that a single static policy will fail to anticipate all the outcomes of the future, yet, over the course of time, uncertainties can be resolved with new information. The adaptive policy-making process is divided into two phases: the "thinking phase," and the "implementation phase" (Walker et al., 2001). The thinking phase is composed of the analysis and deliberation that culminates into an initial policy – essentially, it is the policy formation phase. While all policymaking processes likely involve analysis and deliberation, a planned adaptation thinking phase also includes explicit provisions to enable future adaptation. Part of the adaptive analysis includes identifying metrics, or "signposts," that can be evaluated to judge a policy's performance in order to diminish uncertainty and improve future policies (Walker et al., 2001). Once the initial policy is finalized and rules for execution and evaluation are set, the adaptive process moves into the implementation phase. In this phase, regulators actively monitor the established signposts and metrics to purposely collect information to reduce uncertainty. If collected information "triggers" certain criteria, or if a specific time limit is reached, the policy is reevaluated and adapted to meet the needs of current and predicted future states. For the adaptive implementation phase to work, the thinking phase must not merely state a desire for collecting information and purposeful adaptation, but must also establish formalized pathways and processing to ensure that data collection, analysis, and adaptation can, and will, occur (Eichler et al., 2012).

A single policy can be optimized for a specific set of future states, but may prove brittle if any deviations from assumed futures arise. In fact, many policies are forced to adapt due to unanticipated future states, though this is often accomplished on an *ad hoc* basis (McCray et al., 2009). The view of planned adaptation, in contrast, seeks to anticipate and explicitly plan for adaptation, so it can be accomplished in an efficient and effective matter. In this view, adaptation is not a *reaction* to changes, but rather a *planned* and *expected* part of the policy-making process. This framework, then, recognizes inevitable adaptations as a assumed part of a policy's lifecycle, and not as an unfortunate result of suboptimal decision-making.

7.5 Systems Engineering Digital Engineering Fundamentals

A pre-policy guidance document, released in 2016 by the DoD Digital Engineering Working Group (DEWG), titled "Systems Engineering Digital Engineering Fundamentals (Including Models and Simulations)," offers a relevant example of top level guidance applied to model-centric engineering. It specifies that "[p]rograms should identify and maintain model-centric technology," and that "models and simulations should be used, to the greatest extent feasible." Such a statement makes clear the position that models and model-centric technology should be pursued and used within DoD programs. While this view may spur on development and use of model-centric technologies, it is important to consider the negative implications it might similarly engender. One danger is that it could incentivize susceptibility to "ironies of

modeling" where programs inappropriately model merely for the sake of modeling, to the detriment of effective resource allocation and decision-making. Another point of the guidance offered by the document, however, in line with a general theme of this thesis, emphasizes that all "models, simulations, tools, methodology, and data" used in program decision-making "should have an established level of trust." The reference to trust acknowledges the importance for cultivating appropriate model trust in individuals within a model-centric environment. The next point of the document builds on this by saying that programs "should ensure sufficient training in the appropriate use of models, simulations, tools, data, and the engineering environment." The heuristics developed in this thesis have potential to offer value to both the goals of training and trust development by offering descriptive and prescriptive encapsulations of effective human-model interaction and model-centric decision-making.

Planned adaption for model-centric policies and practices could offer a means to actively collect, analyze, and distribute new information for the betterment of the DoD as a whole. The "Systems Engineering Digital Engineering Fundamentals" states that programs should identify metrics to show how specific training translates into "appropriate use of activities that result in benefits to the program." In addition to identifying effective training as a goal, this statement also rightly acknowledges uncertainty in how best to achieve this outcome. Rather than simply accept this uncertainty, the guidance establishes the need to create metrics to generate insights that will improve training. This same mindset could also be applied to a higher level view of model-centric practices within the DoD in general. With such a mindset, policymakers responsible for the overall DoD vision of model-centric engineering should identify metrics and establish pathways for gathering, evaluating, and distributing new insight gained from these metrics. This information could then be used to create smarter and more effective policies to combat potential challenges to effective model-centric engineering. If such a vision for learning and adapting is to succeed, however, it requires an explicit commitment that moves past dialogue and on to active steps that establish the pathways required for success.

Model-centric capabilities have been identified as desirable by policymakers. Guidance documents, such as the "Systems Engineering Digital Engineering Fundamentals," provide a mechanism for encouraging advancement of model-centric development and use. While expressing clear goals for model-centric use, the "Fundamentals" guidance rightly does not specify *how* to achieve those goals. Overly specific policies and standards, especially early on in a technology's period of innovation and development, can create technical lock-in that stifles creativity and innovation. While new technologies may not need specific policies on how to achieve their goals, relevant empirical evidence is useful for identifying potential challenges and mitigation strategies to guide development. The gathering and analysis of empirical evidence gained through intentionally monitoring the development and use of new technologies or capabilities is an important piece of learning and reducing uncertainty. Initial policies should not be overly specific due to future uncertainty; however, empirical evidence can be used to cut through uncertainty and devise smarter and more effective policies to guide further development and use. In addition to providing guidance for establishing the vision for model-centric practices, policies should also establish pathways to collect and analyze new information to further reduce uncertainty in this field. As demonstrated through the organizational failures that contributed to Chernobyl, organizational learning should be actively pursued, and new insights must also be effectively communicated to all relevant actors who might benefit from the information. Individual areas within of the DoD could innovate and gain insight into challenges, mitigations, or best practices for model-centric engineering, but without established pathways, these insights may not be effectively passed through the whole Department.

8 Conclusions

As this thesis nears the end, it seems fitting to return to the beginning. Exciting things are happening in the world of modeling: we live in a time where technologies like computers, artificial intelligence, and the Internet are allowing us to understand and predict phenomena at levels unprecedented in human history. As these technologies grow and evolve, so do their applications for modeling and decision-making. One manifestation of this evolution is a shift within the engineering world towards model-centric engineering – a paradigm of increasing model integration and use in the engineering process. While model-centric practices are accelerating, this thesis considers a non-technical, yet crucially important factor in the success of model-centric engineering: humans. Although technology evolves and changes with time, people remain the primary benefactor and stakeholder in technology development and use. Any effort to transform the status quo of modeling and decision-making under a digital paradigm must also rightfully consider the intricacies of human-model interaction. The research questions of this thesis aim to prompt research that advances the understanding of human-model interaction in several key areas. This conclusion revisits the research questions and presents a summary of the research that addresses them.

8.1 Research Question 1

- **What human-model interaction challenges exist for individuals placed within model-centric environments, and how might they be mitigated?**

Empirical evidence provides a powerful means for understanding and explaining various phenomena; however, new ideas and technologies, like model-centric engineering, have little data available for empirical analysis. While little empirical evidence exists for human-model interaction in strictly model-centric environments per se, sufficient data and evidence exists concerning human interaction with complex decision-making environments in which they make decisions based on abstracted information. The case studies of pilots interacting with glass cockpits and nuclear reactor operators working within control rooms provide insight into various interaction challenges that influence effective decision-making in immersive, abstracted decision-making situations. Literature on human cognition and decision-making adds additional depth to the understanding of challenges that may hinder effective model-centric decision-making. Two broad categories of challenges to model-centric decision-making emerge: individual challenges and organizational challenges.

Individual Challenges

The commercial airline industry went through a transition from traditional dial instruments to more complex and automated glass cockpits analogous to the transition to a more technical and integrated means of decision-making in model-centric engineering. Additionally, much as model-centric engineering envisions the use of automation to handle the integration of models and information, the glass cockpit introduced computers and automation into the pilot's workspace. This transition revealed various cognitive and perceptual challenges that occur when humans are placed in complex automated environments where they are still the ultimate decision-makers.

As automation began taking over the traditional "stick-and-rudder" role of piloting, a pilot's success has become based on maintaining coherence – "logical consistency in diagnoses, judgements, or decisions" (Mosier et al., 2001). Analysis of three accidents involving glass cockpit equipped aircraft in the 1990's reveal three distinct ways that poor human-glass cockpit interaction can break down coherence: automation bias, automation-induced complacency, and mode error. Automation bias occurs when individuals use automation as a "heuristic replacement for vigilant information seeking and processing," which can lead to errors of commission and omission (incorrectly following an unverified automation directive, and failing

to identify an issue not identified by an autonomous system) (Mosier and Skitka, 1999). Automation-induced complacency occurs when individuals assume a satisfactory system state and fail to properly monitor automation, resulting in poorer detection of system errors and malfunctions than if the system was under manual control (Parasuraman and Manzey, 2010). Mode error represents another pathway to coherence breakdown, and occurs when individuals are unaware of the actual mode a system is operating in. This can cause "automation surprises" where the operator does not know why a system is doing something.

Of primary importance is realizing that automation changes an individual's interaction with a system, and that such a change may have unintended consequences. Systems engineers need to apply user-centered design to create a system that satisfies users' needs, rather than forcing users to adapt to automation. As the glass cockpit case study demonstrates, forcing pilots to adapt to improperly designed or implemented automation creates environments where humans are more susceptible to coherence breakdown, such as automation bias or mode error. User-centered design must apply to what users *need* and not necessarily what they *want*. Preference-performance dissociation, as revealed in the glass cockpit case study, occurs when individuals prefer one system over another, even though they perform better in the other. Individuals who interact with automation must have a strong sense of *accountability* for making correct decisions to mitigate automation bias or complacency. Furthermore, systems need to be appropriately transparent to allow an accurate awareness of a system's state to prevent mode error. Transparency may need to be tailored to specific individuals and tasks at hand, however. If a system is too transparent a user may be overloaded with information; too little transparency and the user may not be able to discern the actual mode in which the system is operating.

Decision-Making and Biases

The dual-process theory of cognition consists of a model composed of two systems for understanding information processing and decision-making, System 1 and System 2 (Kahneman, 2011). System 1 automatically generates emotions and impressions that often influence System 2's analytical and reasoned processing. When faced with non-intuitive problems, System 1 may "help" System 2 by using heuristics to answer simpler problems in an efficient and often effective manner. When heuristics fail to fit a specific situation, however, they create predictable deviations of judgment known as biases. Because these biases happen through System 1's autonomous processing, individuals are often unaware of making biased decisions.

Even with education and training, people tend to be poor at recognizing and mitigating their own biases because they occur automatically and without conscious processing. The key to mitigating biases involves utilizing System 2 processing to consciously and analytically search for a good answer. Organizational structure and policies can force System 2 use in areas that involve a high risk for biased decisions. Checklist use represents one method of forcing individuals to allocate conscious attention to following predetermined normative procedures. Non-advocate reviews present another means for catching and mitigating potentially biased decisions in others. While we may be poor at recognizing our own biases, people are much more capable at identifying biases in others, and reviews that ask proper questions can serve a useful means for revealing biases that might otherwise remain hidden. Activities that require System 2 use, such as following checklists and conducting reviews of decisions, are more resource and time intensive than heuristic-based decision-making, however. Therefore, specific mitigation plans that slow down decision-making should only be in place if the benefits outweigh the costs of the additional attentional and time demands they require.

Model-centric engineering does not aim to provide a system that makes decisions for humans, but rather to improve decisions by helping individuals develop a strongly coherent understanding, or mental model, of a situation. The research on pilot-glass cockpit interaction along with research on individual decision-making shows that improper human-model interaction interferes with this goal by disrupting coherence. The specific means of coherence disruption presented in this thesis are not all encompassing, or even necessarily predictive of errors that may occur with human-model interaction within model-centric decision-making, but they indicate the need for understanding how and why coherence breakdowns occur. Further model-centric research could benefit from an increased understanding of how coherence breakdowns occur in situations specific to model-centric environments.

Organizational Challenges

While nuclear reactor operators, as individuals, can make errors based on abstracted information within immersive control rooms, Chernobyl and Three Mile Island (TMI) highlight the significant influence that organizational factures can play in system failures. These example provide numerous lessons relevant to the development and use of model-centric engineering within organizations. First, certain aspects of poor design can be mitigated upfront by incorporating appropriate human factors considerations into the design. No system design will be perfect, however, so organizations should also have a strong safety culture that actively seeks out flaws and subsequently finds ways to correct them. If issues are discovered, communicatory pathways should exist to spread new, pertinent knowledge to all relevant individuals within an organization. This all points to the need for organizational willingness to learn, fix mistakes, and widely disseminate new findings and improvements. Organizations should also ensure that individuals are properly trained and equipped with the right knowledge and procedures to effectively and safely make decisions and operate a system. These individuals must then be held to a high standard of safety and competence that does not tolerate subpar, unsafe behavior. Both accidents at Chernobyl and TMI involved known defective designs, but this knowledge was not appropriately addressed or shared through the whole organizational system. While the reactor operators in these accidents should not be absolved of all responsibility for their decisions, these organizational factors set them up for failure. Organizations must do their part to ensure that individuals are properly set up for success, but part of that success includes ensuring they are the right people to be handling the information in the first place. If they are properly qualified, the organization should also have a culture that holds them to high performance standards and does not accept inappropriate deviance from established policies and procedures.

8.2 Research Question 2

- **What technological and social factors exist that influence the trust and use of models in a decision-making process?**

The interview-based study performed in this research makes clear that the issues of trust and model use within a model-centric decision-making context are sociotechnical. Trust is an important factor in determining model use: if decision-makers do not trust a model, they will not use it. Various factors, however, can influence an individual's trust in a model. Some of these factors are technical, involving questions about the model itself, while others are more social, involving relationships with other individuals. Figure 13 presents a schematic illustrating that both technical and social factors influence one's trust in a model. Of equal importance is the understanding that these factors differ among individuals. Nevertheless, general trends appear depending on where an individual lies within the decision-making flow of information. Those that have the greatest interaction with models, like modelers and analysts, tend to rely more heavily on technical factors for developing trust. As distance from the model increases, other

actors, like senior decision-makers, tend to place greater reliance upon social factors for developing their trust in a model.

Model trust is an important factor when considering human-model interaction in that it leads to appropriate or inappropriate use of models. All individuals have their own mental model of a situation and of how a model may or may not apply to that specific situation. Individuals may over-trust a model and use it outside its inherent bounds and limitations; conversely, they may also under-trust models and fail to utilize the value they could provide. Calibration involves developing an accurate understanding of a model's capabilities and limitations in relation to a specific decision. A primary goal of human-model interaction, therefore, is to enable proper calibration of an individual's trust in a model. Proper trust calibration requires an understanding of the factors that influence an individual's model trust. Mental model calibration can be affected by social factors, which makes it even more important that everyone within a decision-making flow has a properly calibrated understanding and trust of relevant models.

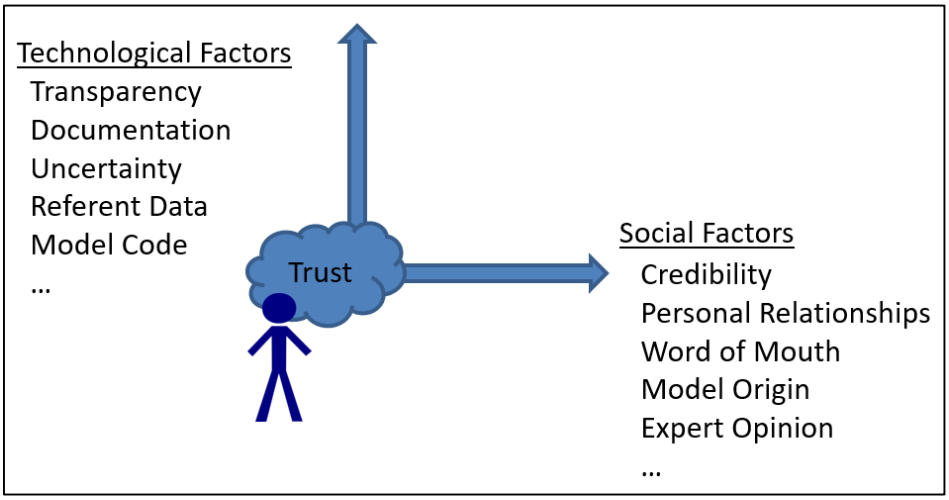


Figure 13. Sociotechnical Factors Influencing Model Trust

8.3 Research Question 3

- **What considerations are relevant for the incorporation of artificial intelligence (AI) and autonomy within model-centric engineering?**

Artificial intelligence (AI) and autonomy are terms prone to misconception and misuse. In many conversations they may be used to imply a complete separation from human control, but nothing is truly autonomous. Artificial intelligence presents a means for enabling autonomy, which itself is a capability that humans can use to automate tasks previously under direct human control. Automation does not displace the task from human control, however, but rather displaces the control in space and time, first to the programmers who create the logic that governs the automation, and second to the operator who chooses how to use the autonomous capabilities in a specific situation. Within this line of thinking, the goal is not to either get rid of human intelligence or augment human intelligence, but rather to extend the intelligence (EI) of the entire system. To do so, AI may displace direct human control in certain areas, but in a manner that extends the total intelligence of both humans and computers operating together. A system that seeks to completely remove all human involvement and control is pointless, but one that stubbornly refuses to replace individuals in certain areas may prove inefficient and ineffective. Therefore, AI must be incorporated within model-centric engineering in a balanced manner, one that augments human intelligence

but may also replace direct human control in certain areas for the purpose of extending the overall system intelligence.

To understand an autonomous system's behavior, one must understand the mental model of the programmers who developed its capabilities. Just as models have bounds and limitations of applicability, automation contains bounds and limitations of effectiveness. A dual responsibility exists on both the part of the programmers and the operator/user to effectively calibrate an understanding of the bounds of a system's autonomous capabilities. Developers have the responsibility to make their mental models containing the automation's capabilities and limitations explicit and understandable. Users, on the other hand, hold the responsibility to properly know the bounds of the automation to ensure that it is applied appropriately. Essentially, users' mental models need to be calibrated to those of the developers (Figure 14). This process is similar to calibrating model trust within a model-centric decision-making process. Autonomy adds an additional layer of mental model calibration – calibration to the automation's capabilities and limitations in addition to calibration to the model itself.

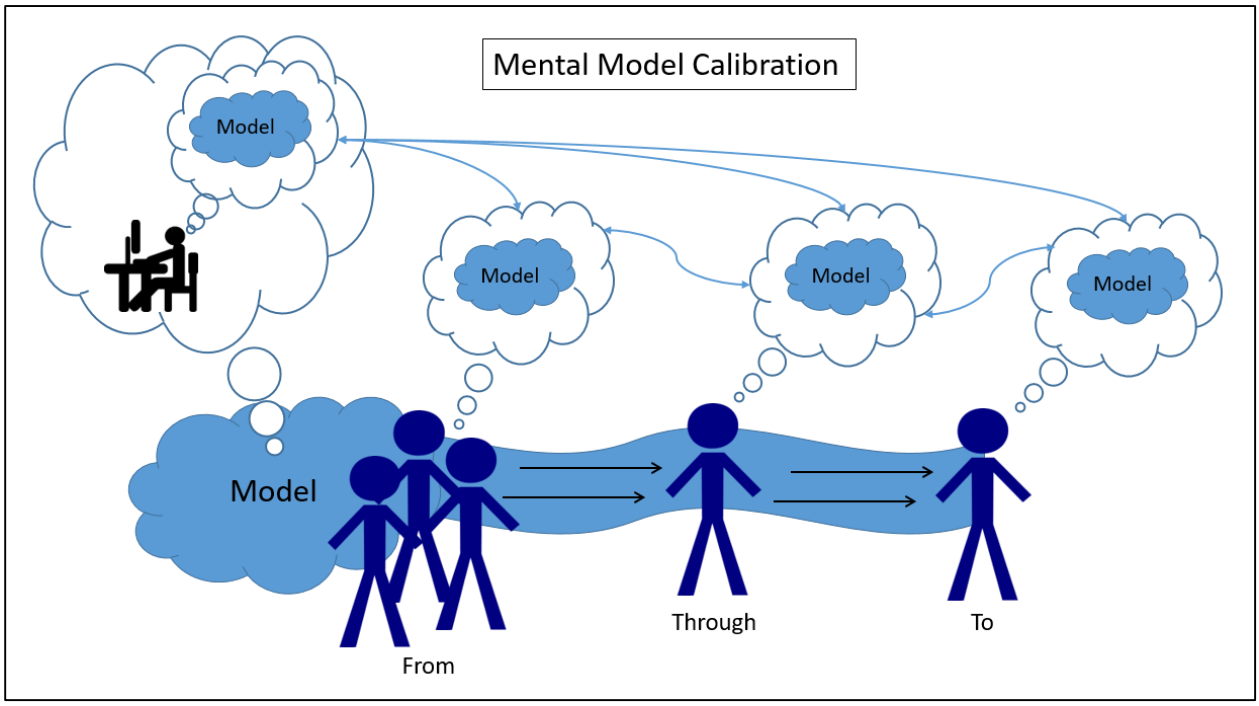


Figure 14. Mental Model Calibration: Model-Informed Decision-Making

8.4 Key Contributions

This thesis presents the following key contributions to the study of human-model interaction and model-centric decision-making:

- **Analogy case studies.** Selected case studies that reveal relevant individual and organizational challenges to effective human-model interaction and decision-making within model-centric environments.
- **Expert interview investigation.** Empirical insight into sociotechnical factors that influence the trust and use of models by various types of actors within the model-centric decision-making process.

- **Considerations for AI and autonomy.** Relevant considerations for how the capabilities of AI and autonomy may relate to a model-centric context.
- **Heuristics**. Descriptive and prescriptive encapsulations of effective human-model interaction and model-centric decision-making to guide future practice, policy, and research. Following further validation, these have the potential to be fundamental principles for use in educating students and the existing workforce.
- **Policy considerations.** Recommendation to pursue a strategy of planned adaption that expands the policy-making process to one with established pathways for reducing uncertainty and iteratively creating more effective policies and guidance for model-centric engineering. Specific policy considerations based on this research are also provided.

8.5 Limitations and Future Research

Much like a model, this thesis has its own limitations. Concerning the analogy case studies, the primary limitation is the generalizability of the analogy findings. While glass cockpits, NPP control rooms, and model-centric environments share similarities in cognitive functions and problem solving through the use of abstracted information between, they are still distinctly different environments that do not offer perfectly transferable lessons. As model-centric practices are accelerating in the world of engineering, future studies could draw empirical evidence from model-centric environments themselves to create more directly applicable lessons.

The size and population of the expert interview study comprises additional limitations. One goal of the study was to understand the perspectives of various actors within the model-centric decision-making process, but not all categories of actors were represented evenly in the sample. The experts in the study came primarily from the *through* and *from* categories, with little direct perspectives from senior decision-makers. Future studies could target individual categories to gain a deeper understanding of their perspectives, which could then be integrated with perspectives from the other categories to form a higher fidelity overall view of the system of actors within the decision-making process. Another limitation stems from the highly qualitative nature of the data and the subjective interpretation needed to generate the key findings. Future research could attempt objective experiments to identify factors that influence trust of individuals. Appendix 2 offers a proof of concept survey experiment for gathering and analyzing data on human-model trust.

The considerations of AI and autonomy primarily rely upon philosophical arguments rather than data. While this does not negate the validity and value of the various assertions made, the arguments do not have the same defensibility enjoyed by evidential arguments grounded in data. Additional, the discussion on AI and autonomy does not give practical guidance on how to specifically implement those capabilities in model-centric environments. Future work could attempt to answer more practical questions such as: How do we balance human, automation, and AI contributions within model-centric environments? What functions do we automate, and how?

Validating the efficacy of qualitative heuristics is challenging in that validation is based on individual opinion, rather than objective data. Members of the MIT research community offered valuable feedback and critiques, but this validation comes from individuals from similar backgrounds and opinions. If time and resources allowed, increasing the pool of expert validators would add greater strength to the quality of the validation results. Additionally, as the discussion on heuristics and biases should have made clear, all heuristics are limited in their descriptive, prescriptive, and predictive power. They should not be viewed as

fully generalizable, universal truths; rather, they offer tools that a skilled craftsman can apply in appropriate contexts.

Similar to the discussion on AI and autonomy, the chapter on planned adaptation of policies and guidance focuses on broader arguments of general theory, rather than practical application. Future policy work could analyze both the thinking and implementation phases of current DoD regulatory practice. An analysis of the thinking phase could assess actual written model-centric policy guidance and regulation to see if adaptive functions are explicitly planned for and incorporated into these policies. Essentially, this "thinking phase" analysis would aim to see if policies align with the normative theory and ideals of planned adaptation. Next, a descriptive analysis of the "implementation phase" could analyze actual programs to see if any sort of adaptive procedures and policies are followed within the organizations, regardless of whether or not they are governed by normative adaptive ideals.

8.6 Final Thoughts

This thesis begins and ends with a call for exploration. Relatively little attention has been paid to understanding human-model interaction; while this thesis hopes to shed some light this problem, its primary aspiration is to motivate further inquiry. Model-centric engineering promises an exciting and innovative future for engineering and decision-making. Innovations, however, bring challenges that can hinder their desired benefits if those challenges are not identified, planned for, and mitigated. The heuristics that emerged from this research aim to provide descriptive and prescriptive encapsulations for mitigating those challenges and enabling effective model-centric engineering. Additionally, an approach of planned adaption for regulation and guidance in this space offers a resilient strategy for identifying prospective challenges and adapting policies effectively.

Model-centric engineering presents a means for helping human decision-makers, not replacing them, with the goal of extending the overall intelligence of organizations. Effective human-model interaction is crucial for enabling the future innovation and success of model-centric technologies and practices. This thesis only begins to scratch the surface of new insights that are waiting to be discovered as we cautiously, yet optimistically and energetically, move towards this new digital future.

9 References

Andre, Anthony D., and Christopher D. Wickens. “When Users Want What's Not Best for Them.” *The Quarterly of Human Factors Applications*, 1995: 10-14.

“Autonomy.” Merriam-Webster. 2016. Web. 12 Dec. 2016.

Bainbridge, Lisa. “Ironies of Automation.” *Automatica*. 1983: vol. 19, no. 6, 775-779.

Baxter, Gordon, Denis Besnard, and Dominic Riley. “Cognitive mismatches in the cockpit: Will they ever be a thing of the past?” *Applied Ergonomics*, 2010: 417-423.

Baron, Jonathan. *Thinking and deciding*. Cambridge University Press, 1988.

Besnard, Denis, and Gordon Baxter. “Cognitive Conflicts in Dynamic Systems.” In Structure for Dependability: Computer-Based Systems from an Interdisciplinary Perspective, 107-124. London: Springer London, 2006.

Blackburn, Mark, Rob Cloutier, Gary Witus, and Eirik Hole. “Transforming System Engineering through Model-Based Systems Engineering (Model-Centric Engineering).” Stevens Institute of Technology, SERC-2014-TR-044-2, Aug. 2014.

Blackburn, Mark, Rob Cloutier, Gary Witus, Eirik Hole, and Mary Bone. “Transforming System Engineering through Model-Centric Engineering.” Stevens Institute of Technology, SERC-2015-TR-044-3, Jan. 2015.

Blackburn, Mark, Mary Bone, and Gary Witus. “Transforming System Engineering through Model-Centric Engineering.” Stevens Institute of Technology, SERC-2015-TR-109, Nov. 2015.

Blackburn et al. “Transforming System Engineering through Model-Centric Engineering.” Stevens Institute of Technology, SERC-2017-TR-101, Jan. 2017.

Bovell, Charles R., Richard J. Carter, and Michael G. Beck. “Nuclear Power Plant Control Room Operator Control and Monitoring Tasks.” Oak Ridge National Laboratory. Mar. 1997.

Boyd, John R. “Organic Design for Command and Control.” Feb. 2005. Lecture. Web. 12 Dec. 2014.

Box, George E.P., and Norman R. Draper. *Empirical Model-Building and Response Surfaces*. John Wiley & Sons, 1987.

Bryant, Antony, and Kathy Charmaz. *The SAGE Handbook of Grounded Theory*. SAGE Publications Ltd., 2007. Print.

Bryant, Antony, and Kathy Charmaz. “Grounded Theory in Historical Perspective: An Epistemological Account.” *The SAGE Handbook of Grounded Theory*. SAGE Publications Ltd., 2007. Print.

Burke, Alafair S. “Neutralizing Cognitive Bias: An Invitation to Prosecutors.” 2 N.Y.U. J.L. & Liberty 512. 2007.

Chappell, Alan R., Edward G. Crowther, Christine M. Mitchell, and T. Govindaraj. “The VNAV Tutor: Addressing a Mode Awareness Difficulty for Pilots of Glass Cockpit Aircraft.” IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans, 1997: 372-385.

“Chernobyl Accident and Its Consequences.” Nuclear Energy Institute. Mar. 2015. Web.

Clifford, Megan M., Mark Blackburn, Dinesh Verma, and Philomena Zimmerman. “Model-Centric Engineering - Insights and Challenges: Primary Takeaways from a Government-Industry Forum.” Stevens Institute of Technology, Jul. 2016.

“Control Rods.” http://www.nuclear-power.net/nuclear-power-plant/control-rods/. Accessed May 12, 2017.

Covan, Eleanor K. “The Discovery of Grounded Theory in Practice: The Legacy of Multiple Mentors.” *The SAGE Handbook of Grounded Theory.* SAGE Publications Ltd., 2007. Print.

Croskerry, Pat. Geeta Singhal, and Silvia Mamede. “Cognitive debiasing 1: origins of bias and theory of debiasing.” *BMJ Quality Safety.* Jul. 2013a.

Croskerry, Pat. Geeta Singhal, and Silvia Mamede. “Cognitive debiasing 2: impediments to and strategies for change.” *BMJ Quality Safety*. Oct. 2013b.

Department of Defense Instruction 5134.16. Aug. 2011.

ODASD(SE). “Digital Engineering.” Systems Engineering, Initiatives, Digital Engineering. Accessed May 11, 2017. http://www.acq.osd.mil/se/initiatives/init_de.html

Eichler, Hans-Georg et al. “Adaptive Licensing: Taking the Next Step in the Evolution of Drug Approval.” *Nature*, March. 2012.

Endsley, Mica R., and Esin O. Kiris. “The Out-of-the-Loop Performance Problem and Level of Control in Automation.” *Human Factors*. Vol. 37, No. 2. Jun. 1995.

Endsley, Mica R. “Automation and Situation Awareness.” Automation and Human Performance: Theory and Applications, 163-181. Mahwah: Lawrence Erlbaum, 1996.

DDRE. “FY011 Department of Defense Human Systems Integration Management Plan.” Washington, DC: DDRE, Director, Mission Assurance, and Director of Human Performance, Training & Biosystems. 2011.

Gass, Saul I., and Lambert S. Joel. “Concepts of Model Confidence.” *Computers and Operations Research*. 8, No. 4, 341-346.

German, E. Shane, and Donna H. Rhodes. “Human-model interactivity: what can be learned from the experience of pilots with the glass cockpit?” 14th Conference on Systems Engineering Research, Huntsville, AL, Mar. 2016.

Gilbert, Daniel T. “How Mental Systems Believe.” *American Psychologist,* Vol. 46, No. 2. Feb. 1991.

Glaser, Barney G., and Anselm L. Strauss. *The Discovery of Grounded Theory*. Chicago: Aldine Publishing Company, 1967. Print.

Hammond, John S., Ralph L. Keeney, and Howard Raiffa. "The Hidden Traps in Decision Making." *Harvard Business Review*, 1998.

Hiremath, Vishal, Robert W. Proctor, Richard O. Fanjoy, Robert G. Feyen, and John P. Young. "Comparison of Pilot Recovery and Response Times in Two Types of Cockpits." In Human Interface and the Management of Information, 766-775. San Diego: Springer Berlin Heidelberg, 2009.

Hutchins, Edwin. *Cognition in the Wild*. MIT Press: 1995a. Print.

Hutchins, Edwin. "How a Cockpit Remembers Its Speeds." *Cognitive Science*, 1995b: 265-288.

"INSAG-7: The Chernobyl Accident: Updating of INSAG-1." International Nuclear Safety Advisory Group. Report. 1992.

"Introduction of Glass Cockpit Avionics into Light Aircraft." National Transportation Safety Board. March 9, 2010.

Ito, Joichi. "Extended Intelligence." PubPub, 1 Mar. 2016. Web. 14 Dec. 2016.

Kahneman, Daniel. Thinking, Fast and Slow. Farrar, Straus and Giroux. 2013. Print.

Kahneman, Daniel, Dan Lovallo, and Olivier Sibony. "Before You Make That Big Decision…" *Harvard Business Review*. Jun. 2011.

Kemeny, John G., et al. "Report of The President's Commission On: The Accident at Three Mile Island." Oct. 1979.

Kumar, Ranjit. *Research Methodology: a step-by-step guide for beginners*. (3rd ed.). London: SAGE. 2011. Print. 2011.

Le Bot, Pierre. "Human reliability data, human error and accident models – illustration through the Three Mile Island accident analysis." Reliability Engineering and System Safety. 2004.

Lee, John D., and Katrina A. See. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors*, 2004: vol. 46, no. 1, 50-80.

"Lessons From the 1979 Accident at Three Mile Island." Nuclear Energy Institute. Oct. 2014.

Markoff, John. *Machines of Loving Grace: The Quest for Common Ground Between Humans and Robots*. HarperCollins, 2015. Book.

McCarthy, John. "What is Artificial Intelligence." Stanford University, 12 Nov. 2007. Web. 12 Dec. 2016.

McCray, Lawrence E., Kenneth A. Oye, and Arthur C. Petersen. “Planned adaptation in risk regulation: An initial survey of US environmental, health, and safety regulation.” *Technological Forecasting and Social Change*. Vol. 77, No. 6. 2010.

Milkman, Katherine L., Dolly Chugh, and Max H. Bazerman. “How Can Decision Making Be Improved?” *Perspectives on Psychological Science*. Jul. 2008.

Mindell, David A. *Our Robots, Ourselves: Robotics and the Myths of Autonomy*. Penguin Random House LLC, 2015. Book.

Minsky, Marvin. “Steps Toward Artificial Intelligence.” Proceedings of the IRE, Jan. 1960.

“Modeling and Simulation (M&S) Glossary.” Department of Defense. Oct. 2011.

Molloy, Robert, and Raja Parasuraman. “Monitoring an Automated System for a Single Failure: Vigilance and Task Complexity Effects.” *Human Factors*. Vol. 38, No. 2. Jun. 1996.

Mosier, Kathleen L., and Linda J. Skitka. “Automation Use and Automation Bias.” Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Santa Monica: Human Factors and Ergonomics Society, 1999. 344-348.

Mosier, Kathleen L., Linda J. Skitka, Melisa Dunbar, and Lori McDonnell. “Aircrews and Automation Bias: The Advantages of Teamwork?” *The International Journal of Aviation Psychology*, vol. 11, no. 1 (2001): 1-14.

Mosier, Kathleen L., Nikita Sethi, Shane McCauley, Len Khoo, and Judith M. Orasanu. “What You Don't Know Can Hurt You: Factors Impacting Diagnosis in the Automated Cockpit.” *Human Factors*, 2007: 300-317.

Mosier, Kathleen L., Linda J. Skitka, Susan Heers, and Mark D. Burdick. “Automation Bias: Decision Making and Performance in High-Tech Cockpits.” The International Journal of Aviation Psychology, 2009: 47-63.

Mumaw, R. J., D. Swatzler, E. M. Rother, and W. A. Thomas. “Cognitive Skill Training for Nuclear Power Plant Operational Decision Making.” U.S. Nuclear Regulatory Commission. Jun. 1994.

Mumaw, Randall J., Emilie M. Roth, Kim J. Vicente, and Catherine M. Burns. “There Is More to Monitoring a Nuclear Power Plant than Meets the Eye.” *Human Factors*. Vol. 42, No. 1. 2000.

Myers, Andrew. “Stanford’s John McCarty, seminal figure of artificial intelligence, dies at 84.” Stanford News. Stanford University, 25 Oct. 2011. Web. 14 Dec. 2016.

Nickerson, Raymond S. “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises.” *Review of General Psychology.* Vol. 2, No. 2. 1998.

Pace, Dale K. “Modeling and Simulation Verification and Validation Challenges.” *Johns Hopkins APL Technical Digest*, 2004.

Parasuraman, Raja, Robert Molloy, and Indramani L. Singh. “Performance Consequences of Automation-Induced "Complacency".” *Aviation Psychology*. Vol. 3, No. 1. 1993.

Parasuraman, Raja, Mustapha Mouloua, and Robert Molloy. “Effects of Adaptive Task Allocation on Monitoring of Automated Systems.” *Human Factors*. Vol. 38, No. 4. Dec. 1996.

Parasuraman, Raja, and Victor Riley. “Humans and Automation: Use, Misuse, Disuse, Abuse.” *Human Factors*. 1997: vol. 39, no. 2, 230-253.

Parasuraman, Raja, and Dietrich H Manzey. “Complacency and Bias in Human Use of Automation: An Attentional Integration.” *Human Factors*, 2010: 381-410.

Pomerleau, Mark. “DoD’s Third Offset Strategy: what man and machine can do together.” Defense Systems 4 May 2016. Web. 27 Oct. 2016.

“Preparing for the Future of Artificial Intelligence.” Executive Office of the President National Science and Technology Council Committee on Technology. 12 Oct. 2016.

Reason, James. “Understanding adverse events: human factors.” *Quality in Health Care*. 1995.

“Research Challenges in Modeling & Simulation for Engineering Complex Systems.” National Science Foundation Workshop Report, January 13-14, 2016.

Rhodes, Donna H., and Adam M. Ross. “Interactive Model-Centric Systems Engineering.” Pathfinder Workshop Report, Feb. 2015

Rhodes, Donna H., and Adam M. Ross. “Interactive Model-Centric Systems Engineering (IMCSE) Phase 3.” Stevens Institute of Technology, SERC-2015-TR-043-1, Mar. 2016a.

Rhodes, Donna H., and Adam M. Ross. “A Vision for Human-Model Interaction in Interactive Model-Centric Systems Engineering.” presented at the INCOSE International Symposium 2016, Edinburgh, Scotland, 2016b.

Rhodes, Donna H., et al. “Interactive Model-Centric Systems Engineering (IMCSE) Phase 4.” Stevens Institute of Technology, SERC-2017-TR-103, Mar. 2017.

Ricci, Nicola, Michael A. Schaffner, Adam M. Ross, Donna H. Rhodes, and Matthew E. Fitzgerald. "Exploring Stakeholder Value Models Via Interactive Visualization." 12th Conference on Systems Engineering Research, Redondo Beach, CA, March 2014.

Rogovin, Mitchell, et al. “Three Mile Island: A Report to the Commissioners and to the Public.” Nuclear Regulatory Commission. Jan. 1980.

Ross, Adam M., and Donna H. Rhodes. “Architecting Systems for Value Robustness: Research Motivations and Progress.” IEEE International Systems Conference. Montreal, 2008.

Sarter, Nadine B., and David D. Woods. “Mode Error in Supervisory Control of Automated Systems.” Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Santa Monica: Human Factors and Ergonomics Society, 1992. 26-29.

Sarter, Nadine B., David D. Woods, and Charles E. Billings. “Automation Surprises.” In *Handbook of Human Factors & Ergonomics*, 2nd Ed, 1926-1943. Hoboken: John Wiley & Sons, 1997.

Shanteau, J. “Competence in experts: The role of task characteristics.” *Organizational Behavior and Human Decision Processes*. 1992.

Skitka, Linda J., Kathleen L. Mosier, and Mark D. Burdick. “Accountability and Automation Bias.” *International Journal of Human-Computer Studies*, 2000: 701-717.

Smith, Marcia S. “NASA’s Space Shuttle Columbia: Synopsis of the Report of the Columbia Accident Investigation Board.” Sep. 2003.

Stanovich, K. E. *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum. 1999.

Stanovich, Keith E., Richard F. West, and Maggie E. Toplak. “Intelligence and Rationality.” In R. Sternberg and S. B. Kaufman (Eds.), *Cambridge handbook of intelligence* (3rd Ed.) (pp. 784-826). Cambridge UK: Cambridge University Press. 2011.

Stanovich, Keith E. “On the Distinction Between Rationality and Intelligence: Implications for Understanding Individual Differences in Reasoning.” *The Oxford Handbook of Thinking and Reasoning*. Nov. 2012.

Stanovich, Keith E., Richard R. West, and Maggie E. Toplak. “Rationality, Intelligence, and the Defining Features of Type 1 and Type 2 Processing.” In J. Sherman, B. Gawronski, and Y. Trope (eds.), Dual processes in social psychology. Guildford Publications, 2014.

Star, Susan L. “Living Grounded Theory: Cognitive and Emotional Forms of Pragmatism.” *The SAGE Handbook of Grounded Theory*. SAGE Publications Ltd., 2007. Print.

Stebbins, Robert A. “Exploratory Research in the Social Sciences.” Sage Publications, Inc., 2001.

Strauss, Anselm, and Juliet Corbin. *Basics of Qualitative Research*. SAGE Publications Inc., 1998. Print.

Strauch, Barry. “Projected Flight Path Displays and Controlled Flight Into Terrain Accidents.” Digital Avionics Systems Conference. Bellevue, WA: IEEE, 1998. E43/1 - E43/8.

“Summer Study on Autonomy.” Defense Science Board. Jun. 2016. Web. 14 Dec. 2016.

“Systems Engineering Digital Engineering Fundamentals (Including Models and Simulations).” Department of Defense Digital Engineering Working Group. Mar. 2016.

“The Role of Autonomy in DoD Systems.” Defense Science Board. Jul. 2012. Web. 12 Dec. 2014.

"Three Mile Island Accident." U.S. Nuclear Regulatory Commission. Backgrounder.

"Three Mile Island: The Most Studied Nuclear Accident In History." General Accounting Office. Report. Sep. 1980.

Tversky, Amos, and Daniel Kahneman. "Judgment Under Uncertainty: Heuristics and Biases." Oregon Research Institute. Aug. 1973.

"Types of Models." in BKCASE Editorial Board. 2017. *The Guide to the Systems Engineering Body of Knowledge (SEBoK),* v. 1.8. R.D. Adcock (EIC). Hoboken, NJ: The Trustees of the Stevens Institute of Technology. Accessed DATE. www.sebokwiki.org. BKCASE is managed and maintained by the Stevens Institute of Technology Systems Engineering Research Center, the International Council on Systems Engineering, and the Institute of Electrical and Electronics Engineers Computer Society.

Vogt, W. Paul, and R. Burke Johnson. *SAGE Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences*. 5th ed. Sage Publications, Inc., 2015.

Walton, Timothy A. "Securing the Third Offset Strategy: Priorities for the Next Secretary of Defense." *Joint Force Quarterly*, Issue 82, 3rd Quarter, 2016. Web. 14 Dec. 2016.

Walker, Warren E., S. Adnan Rahman, and Jonathan Cave. "Adaptive policies, policy analysis, and policy-making." *European Journal of Operational Research*, 2001.

West, Timothy D., and Art Pyster. "Untangling the Digital Thread: The Challenge and Promise of Model-Based Engineering in Defense Acquisition." *INSIGHT* 2015; vol. 18, 45-55.

Wiener, Earl L. "Human Factors of Advanced Technology ("Glass Cockpit") Transport Aircraft." Moffett Field: National Aeronautics and Space Administration, 1989.

Work, Bob. "Remarks by Deputy Secretary Work on Third Offset Strategy." Brussels, Belgium, 28 Apr. 2016. Keynote Speech. Web. 14 Dec. 2016.

Wright, Stephen, and David O'Hare. "Can a glass cockpit display help (or hinder) performance of novices in simulated flight training?" *Applied Ergonomics*, vol. 47 (2015): 292-299.

Zimmerman, Philomena. "MBSE in the Department of Defense." Office of the Deputy Assistant Secretary of Defense for Systems Engineering, US Department of Defense. 2015a.

Zimmerman, Philomena. "A Framework for Developing a Digital System Model Taxonomy." 18th Annual NDIA Systems Engineering Conference, Springfield, VA. Oct. 2015b.

Zimmerman, Philomena. "Advancing Digital Model-Centric Engineering: Digital System Model/Digital Thread." NDIA M&S Committee, Aug. 2015c.

10 Appendices

10.1 COUHES Package: Expert Interview Study

This study was approved by MIT COUHES Protocol 1605567638 and by the DoD OUSD(AT&L) Human Research Protection Official.

	Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects	**Application #** (assigned by COUHES)	
		Date	

APPLICATION FOR APPROVAL TO USE HUMANS AS EXPERIMENTAL SUBJECTS (EXEMPT STATUS FORM)

Please answer every question. Positive answers should be amplified with details. You must mark N/A where the question does not pertain to your application. Any incomplete application will be rejected and returned for completion.

I. BASIC INFORMATION

1. Title of Study	
Human-interaction and decision making within Interactive Model-Centric Systems Engineering (IMCSE) environments	
2. Investigator	
Name: Erling Shane German	Building and Room #:E38-572
Title: Graduate Student	Email: esgerman@mit.edu
Department: Technology and Policy Program	Phone: (719) 428-9064
3. Faculty Sponsor. *If the investigator does not have PI Status (faculty, SRS or PRS) then a faculty sponsor must be identified and sign below.*	
Name: Dr. Donna H. Rhodes	Email: rhodes@mit.edu
Title: Principal Research Scientist	Phone: (617) 324-0473
Affiliation: Sociotechnical Systems Research Center	
4. Collaborating Institutions. *If you are collaborating with another institution(s) then you must obtain approval from that institution's institutional review board, and forward copies of the approval to COUHES).*	
N/A	
5. Funding. *If the research is funded by an outside sponsor, the investigator's department head must sign this form. Please enclose one copy of the research proposal (draft is acceptable) with your application.* ***Do not leave this section blank. If your project is not funded check No Funding.***	
A. Sponsored Project Funding:	

☐ Current Proposal Proposal # _____
Sponsor _____
Title _____

☒ Current Award Account # 6933776
Sponsor Department of Defense (SERC UARC)
Title RT 162 Task Order No. 0062 Interactive Model-Centric Systems Engineering 2016

B. Institutional Funding:

☐ Gift
☐ Departmental Resources
☐ Other (explain) _____
☐ No Funding

6. Statement of Financial Interest

Does the principal investigator or any key personnel involved in the study have any financial interest in the research?
☐ Yes ☒ No
If yes then attach a **Supplement for Disclosure of Financial Interest** for each individual with an interest. *This supplement, together with detailed guidance on this subject and definitions of the highlighted terms, is available on the COUHES web.*

7. Human Subjects Training. *All study personnel in research **MUST** take and pass a training course on human subjects research. MIT has a web-based course that can be accessed from the main menu of the COUHES web site. COUHES may accept proof of training from some other institutions. List the name, MIT or outside affiliation and emails of all study personnel and indicate if they have taken a human subjects training course.*

Erling Shane German, MIT, has completed necessary training
Donna H. Rhodes, MIT, has completed necessary training

8. Anticipated Dates of Research

Start Date: 15 June 2016	Completion Date: 15 May 2017

II. STUDY INFORMATION

1. Purpose of Study. *Please provide a brief statement of the background, nature and reasons for the proposed study. Use non-technical language.*

The Interactive Model-Centric Systems Engineering (IMCSE) project aims to develop transformative results in engineering projects through intense human-model interaction. An invited workshop held in January 2015 seeded a research agenda around the topic of human-model interaction, identifying research needs from both a model-centric perspective and an interactive perspective. Participants agreed that progress has been made on standards, methods and techniques for model-based systems engineering, yet little attention has been given to human-model interaction. IMCSE focuses on human interaction with models and model generated information, and enabling effective model-centric decision making.

As IMCSE is a developing field of research, uncertainty surrounds the nature of the future of IMCSE as well as the aspect of human-interaction and decision making within IMCSE environments. As a means for gaining a better understanding within these areas of uncertainty, this study aims to leverage the opinions of experts and decision makers within the Systems Engineering community. This study will be accomplished through the use of semi-structured interviews that aim to gather expert opinion on specific issues of interest, while also allowing the interviewees latitude to share other experiences and ideas that they also deem relevant to IMCSE and associated human-interaction and decision making with models.

2. Study Protocol. *Please provide an outline of the proposed research. You should provide sufficient information for effective review by non-scientist members of COUHES. Define all abbreviations and use simple words. Unless justification is provided, this part of the application must not exceed 2 pages. Attaching sections of a grant application is* ***not*** *an acceptable substitute for the description requested here. Include copies of any questionnaire or standardized tests you plan to use. If your study involves interviews, submit an outline of the types of questions you will include. Your research outline should include a description of:*

A. Experimental procedures:

Relevant subject matter experts will be contacted through direct email messages or phone calls recruiting them for this study. Willing participants will receive relevant background information, interview questions, and consent forms prior to the beginning of the interviews. Interviews will take place face-to-face if reasonably feasible, or over the phone if that is the more reasonable option. Information from the interviews will be gathered and analyzed for the purposes of furthering the research project's goals.

A sample of the types of questions that will be asked as part of the semi-structed interviews can be found in an attached document titled "Sample of IMCSE Expert Interview Questions."

B. Type and number of subjects involved:

Experts in the field of Systems Engineering to include senior decision makers, program managers, systems engineers, and modelers who may interact with model-centric environments currently, or in the future. Expected 20-30 interview participants.

C. Subject Compensation: (describe all plans to pay subjects in cash or other forms of payment i.e. gift certificate).

No compensation will be provided.

D. Method of recruitment *(attach recruitment materials flyer, poster, email message, Internet posting, etc.)*

Experts will be recruited through direct email messages or phone calls. The attached document titled "IMCSE Expert Interview Recruitment Message" offers a general example of the message that will be contained within the email or phone call.

E. Length of subject involvement:

The interviews are expected to take about 30-45 min.

F. Location of the research:

The interviews are expected to take place face-to-face around the MIT campus or local area for experts who live around the local area and are able to meet face-to-face. Other interviews for experts who are unable to meet face-to-face around the local area will take place over the phone.

G. Procedures for obtaining informed consent *(if a waiver of written informed consent is requested an explanation of an alternative consent mechanism must be submitted):*

The experts to be interviewed will receive a COUHES approved "Consent to Participate in Interview" form prior to the interviews and will respond with appropriate consent before the interviews take place.

H. Procedures to ensure confidentiality:

Unless given permission by the experts to use their names, titles, and / or quotes from the interviews in any publications that may result from this research, the information gathered from the interviews will be confidential.

This project will be completed by 15 May 2017. All interview recordings and transcripts will be stored in a harddrive of a password protected computer located within a locked lab work space until 1 year after that date. The recorded data will then be destroyed.

3. HIPAA Privacy Rule. *If you are in any way working with individually identifiable health information for a research study that is sponsored by MIT Medical, an MIT Health Plan or another healthcare provider, then the Health Insurance Portability and Accountability Act ("HIPAA") likely applies to your study and you must comply with HIPAA in the conduct of your study. However, we expect that if you are applying for exempt status, you will only receive de-identified health information from participants in connection with your study. If you expect to receive identifiable health information from or about research participants in your study, you should complete the standard COUHES application form rather than this application form. You may consult with COUHES staff if you have questions about the exempt/non-exempt status of your proposed research study.*

Signature of Investigator _____ **Date** _____

Signature of Faculty Sponsor _____ **Date** _____

Signature of Department Head _____ **Date** _____

Print Full Name and Title _____

The electronic file should be sent as an attachment to an e-mail: couhes@mit.edu. In addition, two hard copies (one with original signatures) should be sent to the COUHES office: Building E25-Room 143B.

Sample of IMCSE Expert Interview Questions:

- What types of decisions do you make using models?
- How do you judge if a model can be trusted?
- How do you judge if a model is truthful?
- What do you see as advantages of making decisions with models?
- What do you see as risks of making decisions with models?
- What limits your ability to use models to make decisions?
- Do you feel models permit you to make more informed decisions?
- Do you feel models allow you to make more rapid decisions?
- Would you prefer to interact with a model directly, or rather be provided with concise model-generated information?
- How much transparency do you desire (e.g. who created the model, when was it created, what assumptions are inherent in the model) when making decision based off of a model?
- What level of detail to you want to see of the model's analysis when making a decision?
- How desirable would the ability be to directly interact with models real-time while making decisions?
- What types of information do you want to know about a model you are using?
- In the next two years do you expect your use of models in your decision making process to increase, decrease, or stay the same?

IMCSE Expert Interview Recruitment Message

The message that will be contained within the email or phone call will be similar to:

"You have been asked to participate in a research study conducted by Erling Shane German and Donna H. Rhodes from the Systems Engineering Advancement Research Initiative (SEAri) at the Massachusetts Institute of Technology (M.I.T.). The purpose of the study is to leverage expert opinions in the field of Systems Engineering to gain a better understanding of human-interaction and decision making within Interactive Model-Centric Systems Engineering environments. The results of this study will be included in Erling Shane German's Master's thesis. You were selected as a possible participant in this study because of the experience you have as a senior decision maker, program manager, systems engineer, or modeler which we believe may offer valuable insight for this research project. The interview is voluntary and appropriate consent concerning collection and publication of gathered information will be obtained prior to the interviews taking place. Please consider volunteering your time to participate in this project."

CONSENT TO PARTICIPATE IN INTERVIEW

Human-interaction and decision making with Interactive Model-Centric System Engineering (IMCSE) environments

You have been asked to participate in a research study conducted by Erling Shane German and Donna H. Rhodes from the Systems Engineering Advancement Research Initiative (SEAri) at the Massachusetts Institute of Technology (MIT). The purpose of the study is to leverage expert opinions in the field of Systems Engineering to gain a better understanding of human-interaction and decision making within Interactive Model-Centric Systems Engineering environments. The results of this study will be included in Erling Shane German's Master's thesis. You were selected as a possible participant in this study because of the experience you have as a senior decision maker, program manager, systems engineer, or modeler which we believe may offer valuable insight for this research project. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

• This interview is voluntary. You have the right not to answer any question, and to stop the interview at any time or for any reason. We expect that the interview will take about 30-45 minutes.

• You will not be compensated for this interview.

• Unless you give us permission to use your name, title, and / or quote you in any publications that may result from this research, the information you tell us will be confidential.

• We would like to record this interview so that we can use it for reference while proceeding with this study. We will not record this interview without your permission. If you do grant permission for this conversation to be recorded, you have the right to revoke recording permission and/or end the interview at any time.

This project will be completed by 15 May 2017. All interview recordings will be stored in a secure work space until 1 year after that date. The tapes will then be destroyed.

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.

(*Please check all that apply*)

[] I give permission for this interview to be recorded.

[] I give permission for the following information to be included in publications resulting from this study:

[] my name [] my title [] direct quotes from this interview

Name of Subject

Signature of Subject _____ Date _____

Signature of Investigator _____Date _____

Please contact Erling Shane German by email at esgerman@mit.edu or by phone at (719) 428-9064, or Donna H. Rhodes by email at rhodes@mit.edu or by phone at (617) 324-0473 with any questions or concerns.

If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143b, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253-6787.

10.2 Trust in Models Experiment: Final Report

The following report is the final output of a survey experiment accomplished for MIT's class on Human Systems Engineering, taught by MIT professor Leia Stirling. This experiment was approved by MIT's Committee on the Use of Humans as Experimental Subjects (COUHES). It is presented in the Appendix as a proof-of-concept for objectively gathering and analyzing data on human-model trust. Portions of the introduction and literature sections were taken from prior work of the thesis' author. All report co-authors gave approval for the report to be included in this thesis.

The Impact of Information Sources on Human Decision-Making

16.453 Human Systems Engineering
Massachusetts Institute of Technology
Fall 2016

Sarah Folse

Shane German

Eric Hinterman

Kailah Snelgrove

Table of Contents

Abstract

This research study aims to better understand the influence of trust in human decision-making, specifically to determine if decision-makers are more prone to trust computer-generated models versus human information sources. Forty graduate students at the Massachusetts Institute of Technology (MIT) participated in a survey which presented five high-risk scenarios: a Winter Weather Advisory, Blackjack game, Surgical Treatment, Disaster Relief, and Rocket Launch. Each survey question provided a description of the scenario, factors to consider, and advice regarding the decision. The source of advice had three levels of variation: advice given by a computer model, objectively equivalent advice given by a human, and two sources of conflicting advice given by a computer and a human. Each scenario was presented in three variations to encompass these different advice sources, for a total of fifteen survey questions. Additionally, each survey question asked for rationale regarding the decision made by the participant.

Though the scenarios omitted any mention of trust, the word "trust" was used by respondents 68 times when explaining their decisions, clearly indicating that trust is an important factor in influencing decisions. Furthermore, in four out of five scenarios, subjects were significantly more likely to agree with advice when it was provided, showing a higher propensity to trust. However, the results only indicated a statistically significant difference between decisions made when provided advice from a human versus computer in the Winter Weather Advisory scenario; in the other scenarios, no such difference was seen. This study therefore cannot conclusively identify a preference for human or computer advice sources. This is an important finding in itself, which is best summarized by the survey's final debrief question: when given a recommendation course of action, a perfect 50/50 split is observed between those who would trust a human and those who would trust a computer. These results reflect the wide variability in humans' tendencies to trust, and the high complexity involved in decision-making.

Introduction

Background and Motivation

Models represent a powerful tool for informing and influencing decision-making. For models to provide any benefit to the decision-making process, however, human decision-makers must first determine to what degree the models should be relied upon. While various socio-technical consideration likely influence a decision-maker's use of a model, trust seems to be a key factor. The concept of trust has been widely studied in areas such as interpersonal relationships, organizations, and automation, yet research concerning human trust in high risk decision models appears to be lacking. This research project aims to better understand this idea of decision-maker trust in models, specifically to determine if decision-makers place different amounts of trust in computer-generated models versus human information sources.

Fundamentally, models are abstractions of reality that humans use to better make sense of the world and anticipate future events. The quality of abstraction is beneficial in allowing the unmanageable complexities of reality to be broken into understandable and actionable qualities; however, this abstraction also necessarily implies uncertainty in the applicability of a model, as well as various implicit and explicit assumptions that limit the situations for which a model can be considered valid. The oft-cited quote by George Box, "all models are wrong, but some are useful," also implies the importance of determining whether or not a particular model is useful for informing a particular decision. The goal with models is not for decision-makers to use them, but to use them appropriately. If a decision-maker trusts a model too much, this could lead to an over-reliance where the model is inappropriately applied to different situations. Conversely, if a decision-maker has low trust in a model, this could lead to disuse of a model that otherwise could have provided insightful information for a specific decision. Ultimately, the goal for models should be for decision-makers to use them appropriately, and for decision-makers to be engendered with appropriate trust when presented with model generated information.

This study aims to compare how individuals make decisions when presented with similar information from computer models and human experts. The research subjects will be presented with numerous high risk decision-making scenarios that contain information and recommendations from a computer model, human expert, or both. The scenarios will be established such that similar uncertainty surrounds the recommendations from both the model and human expert in order to ensure that the information, though presented from different sources, is objectively the same. The idea is to see if the subjects respond differently to information presented from a computer-generated model versus from a human, with the hope of determining the relative amount of trust subjects place in humans versus models.

Literature

According to Rhodes and Ross, models "can come in a variety of forms and formats, but fundamentally they are an encapsulation of reality that humans use to augment their ability to make sense of the world and anticipate future outcomes." A simulation can be thought of as an execution of a model, where the model provides the "reality" in which the simulation process inputs and produce outputs through the model it executes. This combination of modeling and simulation allows human conceptualizations of a problem space translated to a computer in a manner that produces data and information for the informing of decisions. Although models are increasingly used the decision-making process of various fields from engineering to policy-making, research examining human-model interaction has been lacking. Significant research attention has been paid towards human-machine interaction, however, to include research concerning trust and automation. Automation within systems essentially is a capability afforded through the encoding a models of system response by human programmers. The natural similarities between

automation and modeling and simulation may potentially allow for useful comparisons to be made, and lessons to be shared, between the two fields of research.

Parasuraman and Riley (1997) express the importance of using automation within its appropriate limitations, and describe a taxonomy for human interaction with automation. In their influential article "Humans and Automation: Use, Misuse, Disuse, Abuse," Parasuraman and Riley define misuse "as overreliance on automation (e.g. using it when it should not be used, failing to monitor it effectively), disuse as underutilization of automation, […] and abuse as inappropriate application of automation by designers or managers." While the use, misuse, and disuse of automation involves the individuals specifically interacting with automation, abuse addresses those who design or determine where automation should be used, yet do so inappropriately. This abuse could negatively affect the system, for example, by putting users in situations of inappropriate and unsafe workload, but could also negatively overall performance though failure to train and teach users appropriately.

In their work, "Trust in Automation: Designing for Appropriate Reliance," Lee and See define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." Specifically addressing misuse and disuse, Lee and See express that "[o]vertrust is poor calibration in which trust exceeds system capabilities; with distrust, trust falls short of the automation's capabilities." This idea of calibration "refers to the correspondence between a person's trust in the automation and the automation's capabilities."

Hoff and Bashir build on the idea that trust is an important factor influencing reliance upon automated systems, and identify various factors that may influence trust, and ultimately, reliance upon the automation. Hoff and Bashir divide trust into three components: dispositional, situational, and learned trust. Dispositional trust can be affected by factors such as culture, age, gender, and personality traits. Situational trust is influenced through the internal and external variability that may be present in difference circumstances. Finally, learned trust is gained through learning about the automation along with learning through experience with the automation.

Significant literature, therefore, exists concerning trust in automation, and by extension, machines that use automation. However, the complexity of the topic is nicely captured by a recent article that states, that even following the many years of research on the subject, "the dynamics of trust between people and machines are not yet well-understood." Extrapolating the preceding statement to modeling and simulation, it may also be safe to say that the dynamics of trust between humans and models are not yet well-understood either. Information and insight generated from M&S is often communicated to decision-makers through human actors, which naturally lends itself to influence of interpersonal trust between the humans involved. Although well known as an important factor in interpersonal relationships, perhaps through the anthropomorphizing of technology, or perhaps because it is not a distinctly human phenomenon, trust also is a factor when humans interact with technology. This study seeks to better understand how humans make decision when presented with information from either a human-based source, or a computer/model-based source.

Methods

All experimental data were collected using a Google Forms survey. The survey was distributed to several graduate student list-servs within the Massachusetts Institute of Technology (MIT), and the sample population was therefore comprised of only MIT graduate students. A total of forty subjects participated in this study, all of whom anonymously and voluntarily responded to the survey. The independent variables

in this study include the scenarios (five levels) and the advice sources (three levels). The dependent variables are the subjects' responses to the survey.

The survey required participants to make high-risk decisions in five scenarios: a Winter Weather Advisory, Blackjack game, Surgical Treatment, Disaster Relief, and Rocket Launch. An example question is shown in Figure 1 to the right. The participants were presented with the scenario description, factors to consider, and advice regarding the decision; the participants were then asked to choose one of two multiple choice options and provide rationale regarding their decision. The order in which the multiple choice options appeared was randomized for each subject.

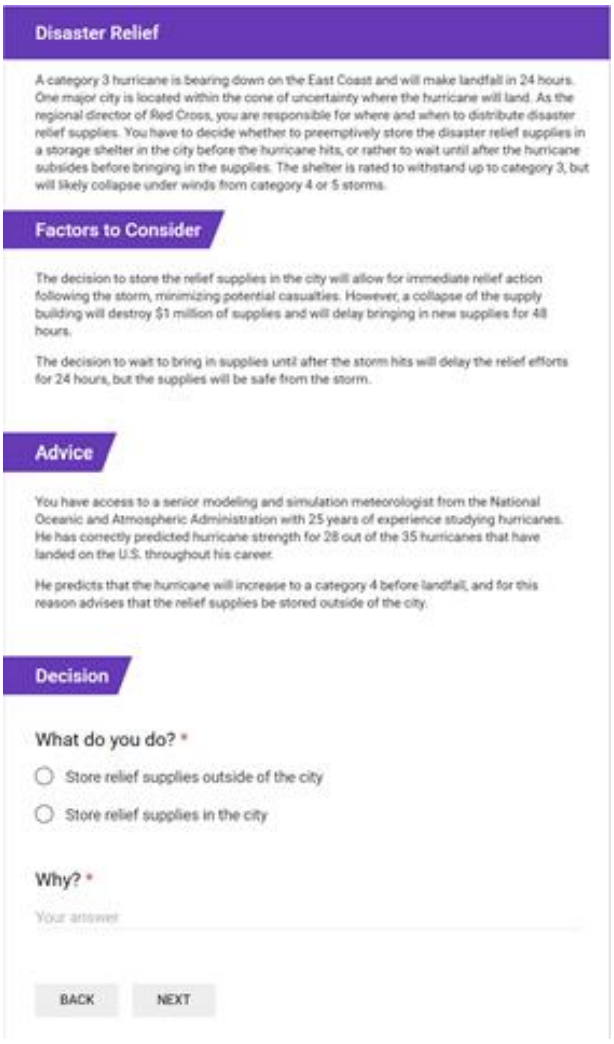


Figure 1: Example scenario from test survey

The source of advice had three levels of variation: advice given by a computer model, objectively equivalent advice given by a human, and two sources of conflicting advice given by a computer and a human. Each scenario was presented in three variations to encompass these different advice sources. Variation A and Variation B presented equivalent information with slightly different wording in the scenario description (i.e., "a 20% probability" vs. "a 1 in 5 chance"); Variation A gave advice from a human, whereas Variation B supplied equivalent advice from a computer. In Variation C, a new scenario description was presented, along with two sources of conflicting advice from a human and a computer (i.e., if the human suggests option 1, then the computer suggests option 2).

Thus, each scenario was repeated three times such that all three levels of advice occurred for each scenario. This resulted in a total of fifteen survey questions. To ensure participants took notice of the variations between scenarios, an "Important Hint" section was shown at the beginning of the survey:

> "Some of the scenarios will seem similar. If this is the case, please ensure that you pay particular attention to the information provided under the "Advice" heading and make your decision using only advice provided on that page. Each page should be treated independently of other pages in the survey."

The order in which the fifteen questions appeared was determined using MATLAB's random number generator, and remained fixed for all participants. At the end of the survey, several debrief questions were included which explicitly asked about the participants' perceptions of trust in both humans and computers.

The results were statistically analyzed using a Chi Square distribution, and all plots of quantitative results were created in Microsoft Excel. Qualitative results were generated using Google Forms.

Quantitative Results and Discussion

After administering the survey to forty participants, results were exported from Google Forms into Microsoft Excel for analysis. Results will first be discussed for each of the scenarios individually, and then an overview of grouped results will be presented. Within each individual scenario, three distinct phases will be discussed that represent the three variations of each scenario faced by the subject: advice from a human, advice from a computer, and conflicting advice from a human and computer. Percentages will be used during the presentation and discussion of results to represent the answers from the forty individuals that participated in the study.

Individual Scenario Results

For the Winter Weather Advisory, 90% of users agreed to enforce a travel ban on the city when presented with this advice from a human, whereas 72.5% agreed to enforce the ban when presented with the same advice from a computer model. When presented with conflicting advice from a human and computer, 62.5% agreed with the human to enforce the ban, while 37.5% agreed with the computer to not enforce the ban. It is interesting that, when presented with the same information twice, eight users switched their initial stance from agreeing with the human to disagreeing with the computer, though each offered equivalent advice. It may be a combination of trust in the human, lack of trust in the computer, and confounding variables that caused this switch to occur. In addition, a higher percentage of subjects decided to not enforce the travel ban when presented with conflicting advice compared to when they were only given advice to enforce the ban (32.5% versus 10% and 27.5%). This result may be linked to a human's tendency to agree with whatever advice is presented; when only given the advice to enforce the ban, people were more likely to do so, whereas when given advice for both scenarios, people were more split in their decisions.

In the Rocket Launch Scenario, 72.5% of users agreed to postpone the launch when given this advice from a human, and 70% of users agreed to postpone the launch when given the same advice from a computer. Only five users total switched in their decisions to launch or not launch the rocket based on the advice they were given, and this was nearly evenly split: three agreed with the computer and disagreed with the human, and two agreed with the human and disagreed with the computer. On the other hand, when presented with conflicting advice from a human expert and a computer model, 60% agreed with the computer to postpone the launch while 40% agreed with the human to launch. The increased amount of split decisions seen in this scenario is similar to that in the Winter Weather Advisory; humans tend to agree with a single source of advice when it is provided but become more varied in their decisions when presented with opposing sources of advice in the same scenario.

With regards to the Surgical Procedure scenario, 100% of users chose to utilize ECMO when given the advice to do so by a computer model, and 95% agreed with a human giving the same advice. The overwhelming response towards ECMO treatment is a potential indicator that this scenario was biased, where users were much more likely to side with ECMO based on the wording. Another possible explanation for the one-sided results is that users are more likely to agree with advice when a human life is at stake, as was the case in this scenario. This interpretation also lends itself to the results of the third section of this scenario, where users were presented with conflicting advice from a computer model and a human expert. In this instance, 87.5% of participants agreed with the computer model to administer HFOV treatment, which could have been considered the more conservative approach by subjects reading the scenario.

In the Disaster Relief scenario, the results were also skewed towards one side for the single advice portions of the situation. The majority of participants followed the advice of both a human and a computer model, with 92.5% agreeing with the former and 95% agreeing with the latter to store supplies outside the city. When the subjects were presented with conflicting advice, only 67.5% agreed with the computer model to store supplies outside the city, whereas 32.5% agreed with the human to store the supplies inside the city. This is another representation of the subjects' tendency to agree with the information source when presented with advice for just one side of the decision, but to be split when presented with conflicting sources of advice.

The last scenario, Blackjack, had more split results than most of those previously discussed. Exactly half of the subjects agreed with the computer model to "hit", while only 32.5% agreed with a human giving the same advice. One of the similar things about the Blackjack case and the Winter Weather Advisory was the high number of people that switched their decision based on what was providing them with advice; eight agreed with the computer but disagreed with the human to "hit", while one agreed with the human but disagreed with the computer to "hit". In the third phase of the Blackjack scenario, participants were given the recommendation to "split" their hands by a computer model and to not split by a human. In this case, 27.5% agreed with the human and 72.5% agreed with the computer. These results may have been influenced by subjects who have past experience with Blackjack and made their own decision and ignored the advice provided. On the other end of the spectrum, the question may have seemed overly confusing for a novice, at which point the subject may have blindly followed advice one way or another.

Grouped Results

Figure 2 below shows the overall agreeableness of the subjects with the advice they were given:

With the exception of the Blackjack scenario, it is important to note that subjects tended to agree with whatever advice was provided to them. It is possible that Blackjack was skewed in the opposite direction because of respondent's familiarity with the game and the given situation.

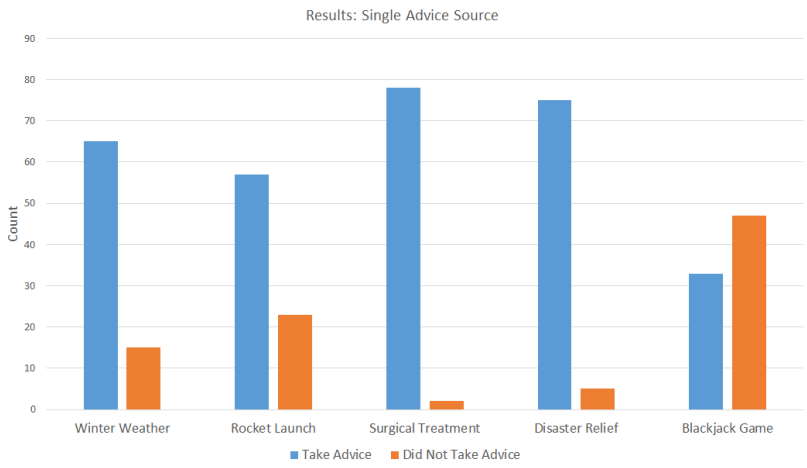


Figure 2: Survey results for each scenario

In order to determine if the advice source matters to the subjects, they were presented with the same advice for the same scenario from both a human and a computer model. The results are shown in Figure 3:

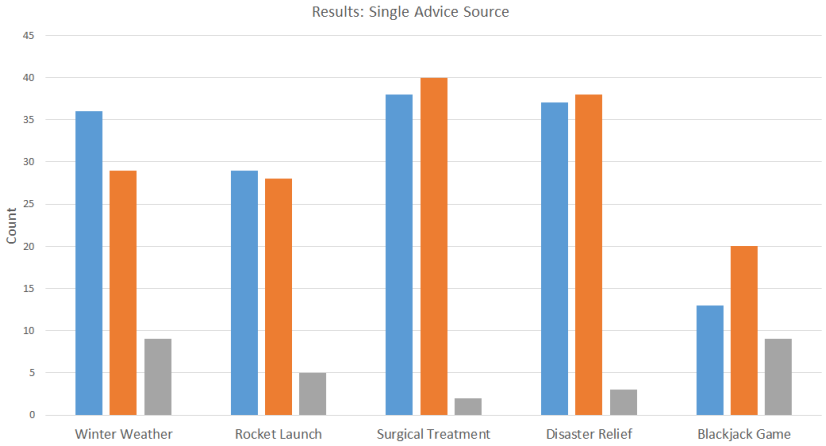


Figure 3: Results for each advice source

As you can see, the Winter Weather and Blackjack game had the largest difference in the number of subjects that agreed with one advice source but not the other. The three remaining scenarios yielded results that were closely aligned between the human and computer model advice.

In order to counter the effects of a human's natural tendency to agree, it is useful to examine the following figure, which shows the results from subjects presented with conflicting sources of advice:

As Figure 4 shows, the results were skewed one way or the other, depending on the scenario. However, when comparing to Figure 2, it can be noted that these results are much closer to being equally split between the two decisions presented to the subjects for each scenario. This indicates that subjects tend to agree with whatever advice is given to them, and therefore when presented with conflicting advice, tend to be more divided with their decisions. These results are context-dependent, but still provide an overall understanding of human cognition in the presence of advice sources.

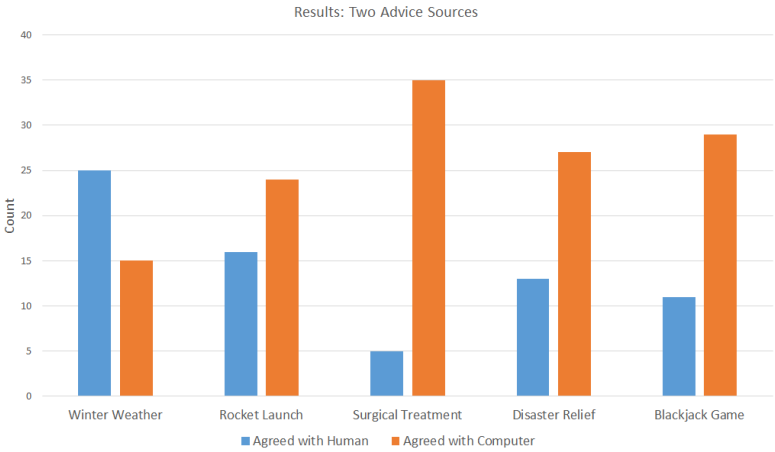
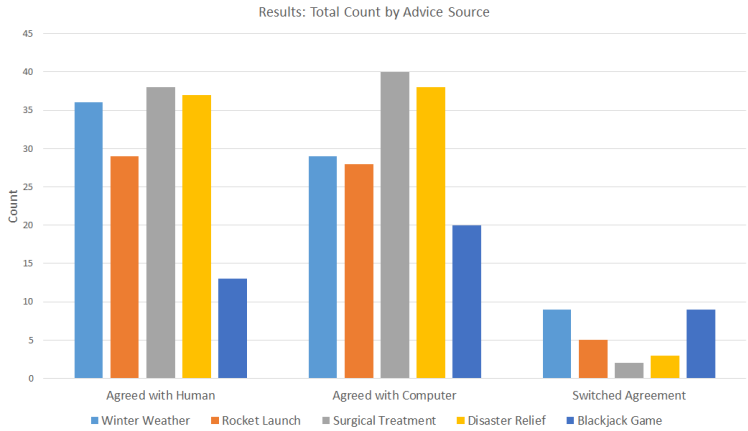


Figure 4: Subject results when presented with conflicting sources of advice

Finally, an overall analysis of the data can help determine if subjects tend to agree more with a computer or a human advice source. This is shown below in Figure 5:



As you can see, subjects had a slight preference for advice given by computers when human life is at stake, as was the case in the Surgical Treatment and Disaster Relief scenarios. Overall, the results are split fairly evenly between agreeing with the computer and human advice.

Figure 5: Summary of results sorted by advice source

Statistical Analysis

In order to analyze the significance of these data, a Chi Square test was employed. This test is used to investigate whether categorical variables with discrete options are statistically different from one another. For this study, subjects chose between two discrete choices for each situation. In the Winter Weather Advisory scenario, they decided whether or not to enforce a winter travel ban on the city in response to a large snowfall. For the Rocket Launch, they gave an order to launch the rocket or to not launch the rocket based on weather conditions. For the Surgical Procedure, they had to decide between administering ECMO or HFOV as a treatment for a sick patient. In the Disaster Relief scenario, subjects chose whether to store supplies inside or outside of the city. Lastly, users decided between "hit" and "stand" for the Blackjack game. In all of these instances, the Chi Square test is appropriate because of the discrete nature of the question. More specifically, a 2x2 Contingency table was used, which is a type of Chi Square test that is defined by the following format:

Variable 2	Data type 1	Data type 2	Totals
Category 1	a	b	a + b
Category 2	c	d	c + d
Total	a + c	b + d	a + b + c + d = N

Figure 6: Chi Square 2x2 Contingency Table

The Chi Square test statistic is then calculated by the following formula:

$$X^2 = N * (ad - bc)^2 / [(a + b)(c + d)(b + d)(a + c)] \quad (1)$$

In this manner, each of the five situations were analyzed for statistical significance across the responses from all forty subjects. Table 1 below shows the results for the Winter Weather Advisory:

Table 3: Example Chi Square results

	Enforce Ban	Don't Enforce	Total
Human Advice	36	4	40
Comp Advice	29	11	40
Total	65	15	80

Ho: The proportion of subjects who enforced the ban is independent of where the advice came from.

Ha: The proportion of subjects who enforced the ban is associated with where the advice came from.

Chi Statistic:	4.02
DOF:	1
Sig. Level	0.05
Chi Critical V:	3.841

Determination: Since chi statistic is outside the critical value, we reject the null.

Conclusion: Therefore, the proportion of subjects who enforce the ban is associated with where the advice came from.

As you can see, the two options the subjects had to choose between were "enforce" and "don't enforce", in reference to the city-wide travel ban. When given advice to enforce the ban from a human expert, thirty-six subjects agreed and enforced the ban, while four chose the opposite. On the other hand, when given the same advice from an equal-fidelity computer model, only twenty-nine subjects enforced the ban. The Chi Square test helps determine if this is a significant difference. The calculated Chi Square test statistic was 4.02, while the critical Chi Square value from the Chi Square table was 3.84 at a significance level of 5% and a calculated degree of freedom of one. Therefore, since the Chi Square test statistic fell outside of the critical region marker, the null hypothesis was rejected. This means we can say that the proportion of subjects who enforced the travel ban is associated with the advice source provided to them.

A similar analysis was conducted on the other four scenarios and found all four to be statistically insignificant compared to the same Chi Square critical value. A table to summarize these results is shown below:

Table 2: Summary of Chi Square test results

Chi Square Test Results: Human Advice vs. Computer Advice			
Scenario	**Chi Statistic**	**Critical Chi Value**	**Reject Null?**
Winter Weather Advisory (Enforce/Don't Enforce)	4.02	3.84	Y
Rocket Launch (Go/No-Go)	0.02	3.84	N
Medical Treatment (ECMO/HFOV)	2.05	3.84	N
Disaster Relief (Store Outside City/Store Inside City)	0.2	3.84	N
Blackjack (Hit/Stand)	2.53	3.84	N

These results lend themselves to interesting analysis. On the one hand, it appears that the source of advice made a significant difference in the Winter Weather Advisory scenario. On the other hand, it did not make a significant difference in the other four scenarios. In considering these conflicting results, it is important to take note of potential confounding variables. For example, some subjects may have found it easier to agree with the human in the Winter Weather Advisory scenario because it would be easier to blame a potential mistake on the human advice-giver than on a computer model that gave advice. In addition, it is possible that personal experiences played a role in shaping people's' decisions, such as familiarity with driving in snow and the dangers associated with it. Overall, these results are interesting and lend themselves to further studies in order to fully separate the dependent variables from the confounding.

Qualitative Results

Our analysis follows two general thrusts. First, the quantitative and statistical analyses seek to understand whether varying the source of advice in a given situation influences the decisions made. This quantitative analysis provides an opportunity to look for patterns and trends in responses; however, this analysis does not conclusively determine causation. The second analytical thrust, therefore, aims to more clearly understand why respondents made the decisions they did. A partial answer to this "why" question results from asking the respondents to provide rationale for their decision in each survey question. The information gleaned from these responses provides insight into individual decisions, and decision-influencing factors may provide a broader understanding of decision-making; however, these data are only directly relevant to individual decision-makers and are not necessarily generalizable. Any generalizations drawn from these results should maintain an understanding of its limitations of descriptive power for populations, as individuals are exceedingly variable in preferences and values, and likewise, will be influenced differently by a wide variety of factors.

Diffusion of Responsibility

One factor involved in subjects' decision-making was a diffusion of responsibility. One individual stated: "I can blame it on the weather man if he's wrong. And kids get a snow day. Everyone loves blaming it on the weather guy." This concept of diffusion of responsibility among humans is a well researched and acknowledged phenomenon. While one subject's response indicated a willingness to diffuse to a human, another subject in a different scenario exhibited the same propensity toward an inhuman source of information: "If I receive backlash I'll offload the responsibility to the computer model." This seems to suggest that diffusion of responsibility may be a factor that influences decisions both when human and non-human actors are involved.

Credibility

The credibility surrounding an individual or a model and the organizations from which they come appears as a salient factor influencing decisions. A brief collection of participant statements relating to the importance of credibility can be seen in Table 3. What exactly comprises an individual's assessment of credibility and how this assessment may vary between individuals promises a rich subject of research that we will not address in this report. Suffice it to say, however, that credibility is a factor that influences a decision-maker's willingness to agree with both human and model sources of information.

Table 3: Statements of credibility from subject's qualitative responses

Influence of Credibility
- "NASA over everything" - "I would trust the experience of the Aerospace Corporation over the model since the Aerospace Corporation has a longer history of success - "NOAA is a legit organization" - "Again, because I trust the institution advising me"

Context

Following the completion of all decision-making scenarios, the survey asked respondents about factors that might influence decisions to either agree with a human source of information, or to agree with a conflicting recommendation from a computer. Numerous subjects indicated that the context of the scenario would prove important in determining which recommendation source to go with. While not providing insight into an individual's propensity for a computer or human leaning bias, per se, these responses do suggest that these individuals are willing to side with humans in certain scenarios and with computer models in others. Table 4 provides a collection of various responses related to the importance of context upon a decision.

Table 4: Responses indicating the importance of scenario context

Importance of Context
- "The context of the decision." - "What type of task is it: one that is easily model-able like a card game or something more dynamic with many more human/social dynamics which would be better for a human recommendation. I know that the computer decision will take into account certain variables so if I know there are variables which are not likely to be accounted for, I will go for a human/my own recommendation." - "The nature of the problem will usually influence my decision. I would be more suspicious of models founded on numerous assumptions, while problems that require a lot of data crunching I would be more suspicious of human recommendations." - "The situation details and what (ex: human lives, money, reputation) is at risk." - "The stakes at play, the robustness of the computer model and its history, and the level of human experience." - "What others will think of the decision, public relations, who designed the computer model, what factors aren't being considered by the computer, how advisers will feel when I take advice from computer instead of them." - "There is likely some technical work/modeling behind a human recommendation, so this seems a false dichotomy. being able to have a conversation with the expert is my main reason."

Trust

Although this study hypothesized, based off of literature and intuition, that trust would prove an important influencing factor in the subjects' decision-making, the word "trust" was intentionally omitted from all survey questions as to not improperly bias participants to think about trust. Despite this intentional exclusion, the word "trust" was mentioned by respondents 68 times in the responses explaining their decisions. These qualitative results clearly suggest that trust is an important factor in influencing decisions. Furthermore, when evaluating the quality of advice given, respondents applied the word "trust" to both humans and computer models, providing support to the notion that trust is not purely a phenomenon found between humans.

In order to gather more information concerning the influence of trust, the survey's debrief questions specifically aimed at understanding how the subjects trust computer-based sources of information compared to human sources of information. Figure 7 and Figure 8 illustrate the results of these queries, which are

nearly identical. An interesting finding is that both of the charts are skewed to the right, indicating a propensity to trust both human and computer sources. The implications of this self-proclaimed trust suggest not only a significant opportunity for sources of advice to influence decisions, but also a tremendous responsibility for both model developers and human advisors afforded by this power of influence.

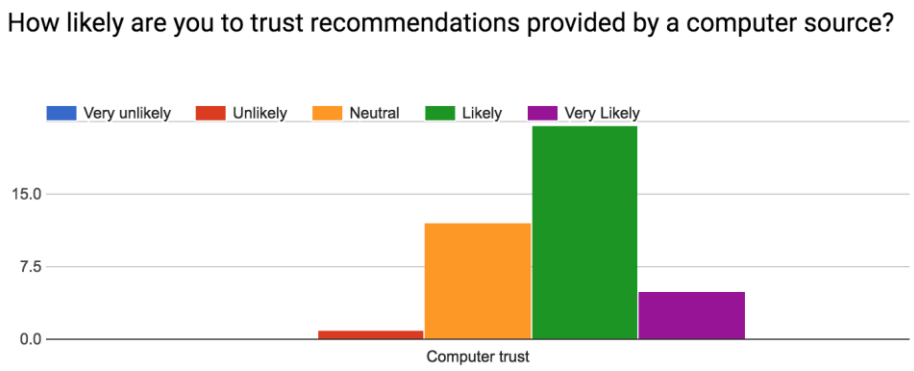


Figure 7: Trusting recommendations from a Computer Source

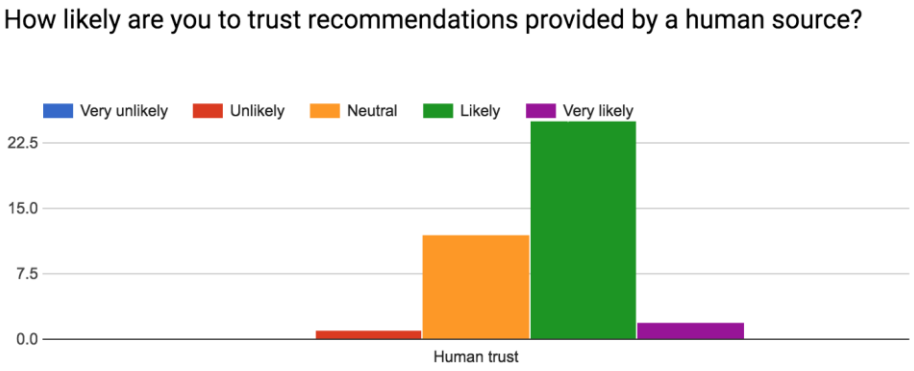


Figure 8: Trusting recommendations from a human source

Although an important factor in decision-making, trust is highly dependent upon the individual making the decision. This point is illustrated by that fact that, when faced with conflicting recommendations, exactly half of the forty respondents believed they would trust the computer, while the other half claimed they would trust the human.

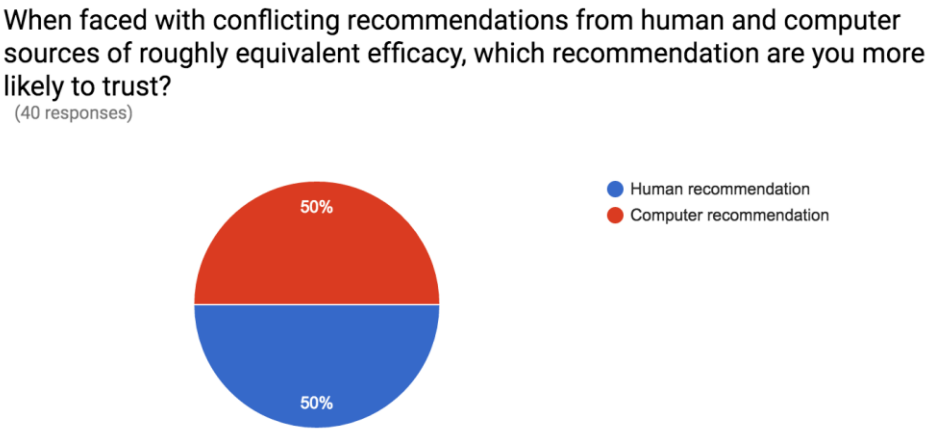


Figure 9: Results when asked to trust either a human or computer advice source

Conclusions

This research study explored the relationships between trust, advice sources, and decision-making using a survey of five high-risk scenarios. Only one scenario - the Winter Weather Advisory - showed a statistically significant propensity to agree with the human over the computer; the other four scenarios did not yield a statistical difference between the two advice sources. Thus, this study cannot conclusively identify a preference for human or computer advice sources, and this is an important finding in itself; the results are best summarized by the final debrief question, in which a perfect 50/50 split is observed between those who would trust a human and those who would trust a computer. These results reflect the wide variability in humans' tendencies to trust, and the high complexity involved in decision-making.

Though defining the role of trust in human decision-making is a complex and multifaceted problem, this study provides strong evidence that it is a vital part of the process, regardless of the entity in which that trust is placed. Additionally, several factors emerged which were outside the scope of this study but played key roles in subjects' responses. These confounding variables included the tendency toward agreement versus disagreement, risk aversion when human life is at stake, and the bias due to organizational branding. Overall, subjects were more likely than not to take advice when a single source was provided, and tended to rank themselves as likely to trust recommendations regardless of the source. In life-endangering scenarios, the data show a preference for agreeing with a human over a computer when one source of advice is provided. Finally, trust in the human or computer was highly dependent on the credentials and affiliation of the source, such as NASA or the Aerospace Corporation.

Future Research

This was an exploratory study, and as such, lends itself to future research in order to follow up on the initial findings presented in this paper. A valuable place to start would be with a duplicate study that reverses the advice provided by each of the humans and computer models. By presenting a new set of subjects with the same set of scenarios but different advice, it could be determined to what extend people's answers reflect the scenarios themselves as opposed to a human's natural tendency to agree with advice regardless of the context. Along the same lines, the same study could be presented with a control group. The control group would not receive any advice, and researchers would be able to therefore establish a baseline to determine if the advice presented to subjects in the test group had an effect on their responses. Both of these additional studies require a similarly large number of participants as this research study (n>30) in order to generate an

effective comparison. Additional situations that could be studied include running the same experiment on a population outside of MIT graduate students, reversing the fidelity levels of the computer and human advice-givers, and generating additional scenarios to complement the five presented in this study. The theme of human versus machine trust and the role it plays in shaping decisions is a difficult area to study. However, it has important implications in the world today, as humans rely more and more on computer models for advice.

References

Darley, J., & Latane, B. (1968). Bystander Intervention in Emergencies: Diffusion of Responsibility. *Journal of Personality and Social Psychology*, *8*(4), 377–383.

DeLaughter, J. (2016). Building better trust between humans and machines. Retrieved December 12, 2016, from http://news.mit.edu/2016/building-better-trust-between-humans-and-machines-0621

German, E. S., & Rhodes, D. H. (2016). Human-model interactivity: what can be learned from the experience of pilots with the glass cockpit? *Conference on Systems Engineering Research*.

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Lee, J., & See, K. (2004). Trust in Automation: Designing for Appropriate Reliance. *University of Iowa*.

Parasuraman, R. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *HUMAN FACTORS*, *39*(2), 230–253.

Ross, A., Rhodes, D., & Grogan, P. (2015). Interactive Model-Centric Systems Engineering (IMCSE).

Tolk, Andreas. The Profession of Modeling and Simulation. *John Wiley.* To be published in 2017.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Zaheer, A., McEvily, B., & Perrone, V. (1998). Does Trust Matter? Exploring the Effects of Inter-Organizational and Inter-Personal Trust on Performance. *Organization Science*.