NATIONAL INTELLIGENCE UNIVERSITY

Ann Caracristi Institute For Intelligence Research

RESEARCH FELLOWSHIP

RESEARCH MONOGRAPH

The Devil You Don't Know

The Need for Joint Human-AI Decisionmaking Outcomes Assessments for Human-in-the-Loop AI Models

Stephen Hood, Ph.D.

Research Fellow, National Intelligence University, 2020-21

The Devil You Don't Know

The Need for Joint Human-AI Decisionmaking Outcomes Assessments for Human-in-the-Loop AI Models

Stephen Hood, Ph.D., National Geospatial-Intelligence Agency

RESEARCH FELLOW, NATIONAL INTELLIGENCE UNIVERSITY, 2020-21

PUBLISHED SUMMER 2023

Abstract

People have a love-hate relationship with artificial intelligence (AI) agents that must be addressed to successfully implement human-AI teams. On the one hand, AI agents have demonstrated the potential to improve the accuracy and speed of human decisionmaking. On the other hand, the agents are known as "black boxes" that produce recommendations based on inputs and processes that are not clear to end users and that have been shown to alter the decisions these end users might otherwise have made. Existing research on this topic focuses on two areas: identifying factors in the human-machine relationship that influence decisionmaking and identifying strategies to improve the joint decisionmaking environment. This NIU Research Monograph uses an empirical approach to explore the need to account for the human element in developing a constructive human-AI relationship. Specifically, this work contributes to the field of human-AI interaction in two ways. First, two new factors influencing human decisionmaking in the AI-human team are identified: *Self-Assessed Expertise* (i.e., the human participants judge themselves as expert or nonexpert on the subject of interaction with the AI agent) and *User Interface Settings* (i.e., the format used by the AI agent to present recommendations). Second, this research aims to provide data-driven recommendations for improving the overall quality of decisionmaking of the human-AI team. Two studies have been conducted, demonstrating that both performance and task engagement improve when people are allowed to customize the level of AI output explainability—i.e., the detail they receive that helps them understand the solutions offered by the AI agent. Finally, implications and recommendations for both managers and researchers are presented that address the implementation of AI agents in analytic settings.

Key Findings and Recommendations

- Based on the author's study of human-AI agent interaction in a controlled setting, self-assessed nonexperts are significantly more likely to accept an AI agent's recommendations (i.e., user interface settings) when they are presented differently each time. A postexperiment survey measuring study participants' reliance on AI agent's recommendations suggests this acceptance is a conscious choice.
- Although experienced professionals are more accurate at cognitive tasks when working with AI agents that offer relatively more explainable recommendations, self-assessed human experts also display only grudging engagement with more sophisticated algorithms. This apparent aversion to the AI agent runs contrary to the literature's Explainable AI (XAI) hypothesis, which states that designing AI agents to provide relatively more interpretable recommendations to their human counterparts improves human-AI interaction.
- Experienced professionals are also more accurate when given a *choice* in explainability level, regardless of the explainability level chosen, and their dislike toward the AI agent declines. Although study participants self-reported greater reliance on the AI agent's advice when receiving more explainable recommendations, they did not, in fact, rely more on the AI agent's guidance. This suggests that allowing people to choose their explainability level may be more important to maximizing joint human-AI accuracy than merely increasing explainability.
- Managers and researchers seeking to introduce AI agents into existing workflows, therefore, should work to optimize both singular AI agent performance (i.e., proportion of false positives or negatives) and the joint performance of the human-AI team. Situational and environmental factors may influence not only end users' acceptance of recommendations, but also the overall performance of end users working in concert with the AI agent.
- Managers and researchers should consider developing sustainable approaches to identifying, measuring, and balancing the effects of individual differences (e.g., expertise level and subsequent reliance on the user interface) across analytic teams, as well as standardizing user interfaces for joint human-AI systems.

Contents

List of Figures

List of Tables

Preface

In its overview of the strategic environment, the 2019 National Intelligence Strategy (NIS) calls out emerging technologies such as "artificial intelligence, automation, and high-performance computing … [as] economically beneficial" and capable of enabling "new and improved military and intelligence capabilities for our adversaries."[1] The use of such technologies offers the U.S. Intelligence Community (IC) significant advantages relative to traditional analytic approaches that are both time and personnel intensive. For example, AI does a better job synthesizing and making decisions based on vast amounts of data. AI is often more precise in its predictions and better able to conceptualize and act on decisions where success depends on statistical reasoning. It is faster.

But, as the 2019 NIS implies, not all that glitters is gold. Successful implementation of AI systems faces key challenges related to development and delivery, ethics, data sharing, and adoption. People tend toward irrational avoidance of or attraction to AI agents (i.e., human-AI reactance), and these behaviors can sometimes result in suboptimal joint human-AI decisionmaking.

While researchers have sought to mitigate these challenges by keeping a "human in the loop," the attraction of low-cost and fast decisionmaking tools has enticed some competitors to increasingly cede decisionmaking authority to AI-powered and autonomous weapon systems. To maintain the United States' edge, U.S. decisionmakers have followed suit, and this pattern all but ensures an ever-shorter portion of the "loop" within which humans maintain control over AI counterparts.[2]

Sustaining human oversight requires AI implementation approaches that identify not only factors influencing the joint human-AI decisionmaking process, but also ways to improve overall decisionmaking outcomes. This work seeks to improve understanding of some key factors and methods of human-AI decisionmaking. First, this monograph provides an overview of the relevant joint human-AI decisionmaking literature and highlights existing knowledge gaps. Second, it reports on the results of two sets of experiments that identify both additional drivers of human-AI reactance and approaches to improving joint human-AI decisionmaking outcomes, adding to the growing body of literature that addresses these issues.

This research is the culmination of a 12-month Research Fellowship at the National Intelligence University's (NIU) Ann Caracristi Institute for Intelligence Research (CIIR). It was made possible through the support of CIIR staff and faculty, the author's research committee, and the 2019-21 CIIR Fellows cohorts, as well as countless offices and individuals at the National Geospatial-Intelligence Agency (NGA) who consulted and collaborated on various elements of this effort.

Background: The Troubled Relationship Between AI and Analysts

During the past several decades, the U.S. IC's mission requirements have outpaced the Community's capacity to satisfy them. This mismatch results from the confluence of several factors including technological improvements facilitating collection of increasing amounts of data, a shifting competitive environment in which the United States must increasingly compete against near-peer nations, and a budget that has not grown proportionate to these demands. This trend is likely to continue, and the IC has increasingly turned to artificial intelligence (AI) systems as one approach to bridge the gap between requirements and capacity.[3, 4]

This approach is understandable. Researchers have long been aware that statistical algorithms, of which modern AI systems are comprised, can outperform their human counterparts in a variety of tasks. Early examples include Virginia Apgar's (1953) proposed algorithm for evaluating the health of newborn infants and later Paul Meehl's (1954) "Disturbing Little Book," in which he showed a variety of other settings in which algorithmic statistical prediction consistently outperformed domain experts' subjective predictions.[5, 6]

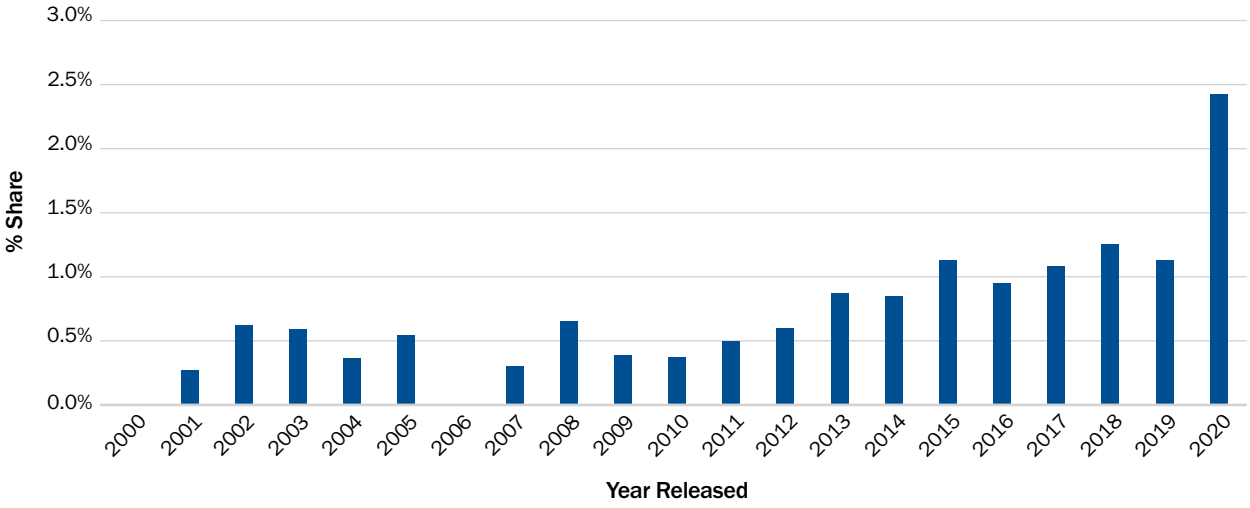
The next 70 years saw an explosion of interest in developing algorithms capable of predicting and classifying increasingly complex data. For example, research supporting current approaches such as neural networks and K-nearest neighbor algorithms began in the 1950s.[7] However, because of the necessary computational intensity and oftentimes tremendous data requirements, many of these algorithms remained of primarily theoretical rather than practical interest.

Recent advances in technology have made available vast amounts of low-cost data, cheap storage, and greater processing power. These trends are expected to continue and have opened the door to investment in AI-driven solutions. For perspective, U.S. Defense Department investment in AI has grown from $600 million in FY2016 to $926 million in FY2020, and private sector investments in 2018 were estimated at nearly $8 billion.[8, 9] However, although early work leveraging AI against real-world problems has yielded some success, the IC has yet to obtain its expected return on investment.

Specifically, conversations with IC analysts and managers who use these algorithms reveal a troubled relationship. Although some analysts willingly accept AI systems augmenting their workflows and even report enjoyment working with them, other analysts seem to dislike working with AI systems and eschew their use

whenever possible. These behaviors are consistent with *algorithm appreciation* and *aversion*, which describe the attraction or antipathy phenomena in which people irrationally overuse or underuse an AI system.[10, 11] This paper refers to such behaviors collectively as AI reactance. Such concerns and reactions are valid. In many instances AI systems have failed completely or have failed in such a way as to accentuate human biases.[12] Left unchecked such AI systems may do a great deal of harm, and evidence of public concern surrounding such potential failures is widespread. For example, consider the increasing market share of movies produced in which AI plays a key plot element (see Figure 1). In 2020, nearly 2.5 percent of all movies produced contained a key AI plot element, roughly double the percent in each of the five previous years.[13] Clearly, this is a topic that concerns the broader public.

Figure 1. Market Share of Films Produced Featuring AI



Author's figure based on the following source: "List of Artificial Intelligence Films," Wikipedia, accessed on September 15, 2021, https://en.wikipedia.org/wiki/List_of_artificial_intelligence_films.

Human-in-the-Loop Solutions

To mitigate some of these concerns, AI system proponents advocate for human-in-the-loop (HITL) designs. HITL refers to the idea that, although AI systems are capable of learning from human behavior and then acting autonomously, a human should oversee or work alongside the AI system and be the ultimate decisionmaker. The challenge with ensuring human oversight is that people frequently embrace algorithm recommendations without realizing it. Consider, for example, autocompletion in Intellipedia's main search bar.

In such instances, some human-in-the-loop algorithms have been programmed to learn from users and adjust their output based on perceived user preferences. Although these algorithms may be designed to "assist" in simple tasks and their features may be transparent to users, they are likely to influence preferences

and decisionmaking in ways that are difficult to predict and even subperceptual.[14, 15] Thus, seemingly innocuous AI system recommendations have the potential to cascade into consequential decisionmaking biases without a user's awareness. This poses a concern in two areas.

First, end users may be susceptible to overacceptance or rejection of AI systems as a function of situational factors or individual differences. The varying degrees of appreciation or aversion that different individuals have for applying AI to their decisionmaking process and other environmental factors—including, as will be seen in the studies reported in this monograph, ways in which an AI agent communicates its recommendations—introduce a potentially unbalanced, stochastic (i.e., random) element that should be accounted for in joint human-AI decisionmaking. In this monograph this type of behavior is referred to as AI reactance.

Second, AI systems may change the decisions made by the end users. For example, any given decisionmaking outcome may be worsened or improved by joint human-AI teaming, making it necessary to understand and further research those factors that can improve decisionmaking.

Monograph Overview

Based on these two possibilities, IC managers and AI systems developers should be concerned with two different aspects of AI system integration into workplace settings. The author presents these in two research questions (RQs) and explores them in the context of the current literature and two sets of experimental surveys.

First (RQ1), how does the introduction of AI systems into the workplace influence human decisionmaking, and can tangible drivers of AI agent reactance be identified?

Second (RQ2), how can joint human-AI decisionmaking outcomes be improved?

The next section presents the research question in terms of the relevant literature, including a brief historical overview and the streams of research that contribute to the current investigation. Specifically, the author provides a short history and overview of algorithms generally and how they relate to modern notions of AI and its subordinate applications. This discussion is followed by a review of popular and academic perspectives on human-AI reactance, how human-in-the-loop designs influence both performance and AI receptivity, and the role that domain expertise plays in these considerations—all framed in the context of a tangible, environmental factor (*User Interface Settings*) and how that relates to end user domain expertise, which is an important consideration in the IC. Second, the author explores the need to identify factors that improve the overall quality of outcomes in joint human-AI decisionmaking. This discussion includes the contributions of the Explainable AI literature and the psychological benefits of *Choice*, which together form the theoretical foundation for the author's second set of experiments.

The third section provides an overview of the methodological approach used to investigate the research questions. The author's research relies on an empirical survey design, and this section discusses generalized versions of both sets of experiments, including the decisionmaking environment and manipulated variables, as well as empirical design choices in the context of the approach used to analyze the resultant data.

The fourth section presents the findings of the author's two sets of experiments. In the first set, the results contribute to the algorithm reactance literature and specifically show that nonexperts are more susceptible to AI agent-provided recommendations when they have *User Interface Settings* that are inconsistent with their previous experience. The second set shows that allowing end users a choice in the explainability level of an AI agent's recommendation (i.e., how much detail is provided on factors that determined the recommendation) not only improves the accuracy of the end user's decision but also improves human engagement with the AI agent.

The fifth section discusses findings from both sets of experiments in terms of the existing literature on algorithm reactance and choice. The experiments' results advance the argument for additional research into choice and for allowing end users to customize choice and additional human-AI touchpoints when participating in human-AI agent teams. The goal for this monograph is to highlight the importance of accounting for the human element in the human-AI relationship when implementing AI agents in IC settings. In particular, this paper seeks to investigate additional approaches to: identify and mitigate drivers of human-AI reactance (RQ1) and improve joint human-AI decisionmaking (RQ2).

In the sixth and final section of the monograph, implications from the results and follow-on discussion are presented along with recommendations that are accessible to IC managers and system developers. The author presents recommendations derived directly from this research, as well as recommendations for research extensions. Finally, the author concludes with the assertion that managers and developers should evaluate the success of AI implementation in terms of joint human-AI outcomes.

Algorithms and AI

This section provides a brief history of algorithms and their contextual relevance to modern notions of AI. The relevant streams of research that inform this paper's overall research question are discussed, and several hypotheses that drive the paper's research objectives are offered.

Algorithms: The Backstory

The algorithms that form the basis for modern artificial intelligence systems have been around for a long time. At their simplest level, algorithms are a set of instructions designed to produce a consistent result. Although they often involve mathematical processes, this need not always be the case. For perspective, the earliest examples of algorithms include a Sumerian system for division that emerged around 2500 BCE and a Babylonian approach to calculating inverses that was developed around 1600 BCE. Their early and widespread appeal was a function of their usefulness. Not only could nonexperts (provided they could read) use them to successfully execute relatively complex sets of instructions that would have been otherwise inaccessible, but they also improved the precision with which people executed those tasks.[16]

The appeal of algorithms extended into the modern era. In the 20th century, increases in wealth and subsequent global demand for progressively technologically advanced goods led to the popularization of mass production and a growing need for the increased precision such algorithms provided.[17] A need for consistency and notions of fairness soon led to widespread applications of algorithms. These applications included such disparate fields as cryptography and law, and shortly thereafter began to evolve into the mechanical and automated computational approaches that are consistent with contemporary notions of algorithms and AI.[18, 19, 20]

Algorithms as Recommendation Agents

The same consistency and precision that increased the appeal of algorithms in accounting, production, and other similarly process-driven disciplines also made them attractive recommendation systems for disciplines previously dominated by expert opinion. Virginia Apgar (1953) is credited with one of the earliest such findings in which she developed an algorithm to systematically assess the health of, and inform subsequent treatment for, infants in the moments immediately following childbirth. Her algorithm was straightforward: physicians would assign a 0, 1, or 2 under five health dimensions (heart rate, respiratory effort, reflex irritability, muscle tone, and color). Infants with higher scores tended to be healthy, whereas

those with lower scores tended to be unhealthy and require emergency intervention.[21] As a testament to its success, versions of this algorithm are still recommended by both the American Academy of Pediatrics and the American College of Obstetricians and Gynecologists' Committee on Obstetric Practice.[22] Apgar's work was later joined by that of Paul Meehl (1954), whose monograph investigating comparisons between algorithmic and expert prediction generalized these findings to a wide variety of situations including assessment of future academic performance, as well as prediction of parole violations and criminal recidivism.[23, 24]

Since then, numerous other examples have been found in which simple algorithmic recommendation systems outperform their expert human counterparts. In 2002, researchers found evidence suggesting that algorithmically generated mortgage underwriting more accurately predicted mortgagee default than human underwriters. Furthermore, use of the algorithmically generated mortgage underwriting system not only generally increased borrower approval rates with lower default risk (resulting in increased revenue for the underwriting firm), but especially increased borrower approval rates for underserved populations resulting in improved social welfare.[25] Similar findings have been found for other domains: AI systems used to screen résumés increased a firm's selection of "nontraditional candidates" relative to human screeners, based on field data; Child Protective Services' algorithms have been credited with doing a better job at identifying at-risk youth than human screeners; and, in an extension of Paul Meehl's work, recent research shows that even simple linear models tend to outperform individual expert judgment.[26, 27, 28]

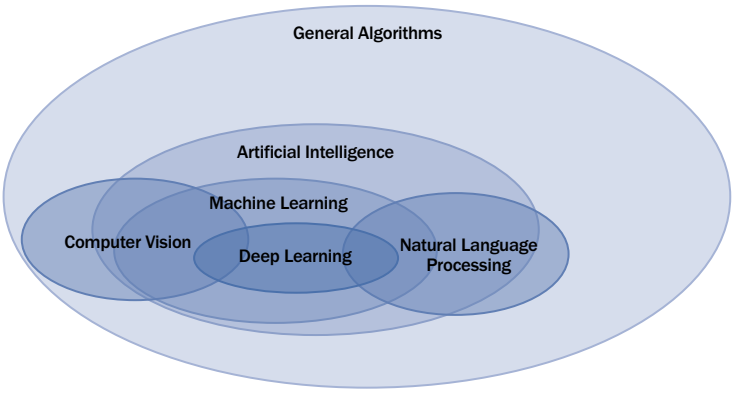
Nevertheless, many workers disliked working with early algorithms. Assembly line workers famously hated Henry Ford's assembly line, preferring instead the challenge of working as a team to assemble a whole car at one station. Reasons cited included reduced satisfaction stemming from performing only one task and never seeing a completed product, monotony, and perceived loss of self-determination.[29] In recent years, research into the fields of operations research, management, human factors, and other related disciplines has identified and mitigated many of the most common complaints against such simple algorithms, and adoption of simple algorithms is now widespread. Examples of successful adoption include contemporary cognates of the earliest algorithms such as multiplication and division tables, as well as more modern computational processes such as internet search and vehicle navigation. At a basic level these are prescriptive calculations of an optimal path. At a more advanced level they are adaptive and offer recommendations based on dynamically changing factors.

Complex Algorithms

The development of complex algorithms proceeded quickly in the 20th century. Many of these advances were enabled by concurrent developments in computing which afforded cheaper data collection, storage, and processing power.[30] Early integration of algorithms with computational approaches (e.g., Alan Turing's "Turing Machine") gave rise to greater interest in the ability of computers to execute increasingly complex sets of instructions that were adaptive to underlying data, and this culminated in the modern notion of AI that allows greater adaptational flexibility.[31] Results from these early successes led to further demand and research into machine learning (ML), deep learning (DL), and other subsets of AI that were designed to accommodate increasingly adaptive clusters of processes depending on the underlying data encountered (see Figure 2).[32]

Thus, many of the "advanced" AI applications such as machine learning (ML), computer vision (CV), and natural language processing (NLP) that drive modern prediction and recommendation systems have much earlier antecedents as statistical algorithms. For example, the K-nearest neighbors algorithm, which is a form of supervised ML, owes its existence to research conducted in the 1950s by Evelyn Fix, Joseph Hodges, and Thomas Cover.[33, 34] And current advances in CV find their origin in a 1966 summer project in which researchers attached a camera to a computer and tried to get the computer to describe its surroundings.[35] This paper uses the term AI to refer to the entire class of adaptive algorithms designed to classify, predict, or provide recommendations.

Figure 2. Relationships Among Algorithms, AI, and Subsets of AI



However, many of these early advances remained primarily of theoretical interest. Early computers were expensive, and few had the processing power necessary to execute more advanced AI algorithms. Furthermore, more sophisticated algorithms required vast amounts of data that were not readily available. Increased proliferation of computer systems beginning in the 1980s led to faster collection and cheaper storage of data, cheaper processing power, and increased demand for practical applications of otherwise esoteric statistical approaches.

Contrary to the reception of simple algorithms, more complex AI algorithms that promised significant benefits to end users did not enjoy the same widespread acceptance. Instead, they became something of an object of fascination (and horror)—rather than expecting to see them at work, people expected to see them in the cinema and on the bookstand, often playing the role of an antagonist, such as in the popular films *The Terminator* and *The Matrix*.

Human-AI Reactance

Instances of aversion to complex algorithms also became an object of fascination to researchers in the social and business sciences. In much the same way that researchers in the early 20th century explored ways to overcome the psychological challenges associated with the assembly line, it became necessary for researchers to investigate how to overcome psychological challenges associated with working with more advanced AI agents.

Notable examples of such research include work by U.S. economist and decision theorist Berkeley Dietvorst who has popularized the term "algorithm aversion," which he defines as an irrational preference for nonalgorithmic advice despite an algorithm's proven effectiveness. Specifically, he has demonstrated that, when participants in an experiment were shown both an algorithm and a human making a mistake, they were more likely to accept the recommendations from the human, and this effect persisted even when the algorithm

outperformed the human. Thus, participants were less forgiving of algorithmic mistakes than human mistakes, and this influenced their subsequent tendency to prefer human over algorithmic forecasts.[36]

Moreover, when given the choice between working with algorithmic vs. human advice, participants in a study undertaken at the University of Bath tended to overweight human advice. In a set of experiments, the researchers asked participants to forecast stock prices based on a provided information set. After the participants made their initial predictions, the researchers gave the participants advice from a human expert as well as a statistical forecasting algorithm, and then asked the participants to reevaluate their initial forecast. Results showed that the participants subsequently overweighted advice from the human expert and underweighted advice from the algorithm. In a separate study, these researchers provided participants with advice from either two experts or two statistical algorithms; in this case, the participants weighted advice equally. Thus, when people had the choice to take advice from a human expert or a machine in a decision-making task, they gave greater value to the human expert than to the machine. However, when not given the opportunity to choose between humans and algorithms, people weighted the advice equally, irrespective of whether it came from a human or algorithm.[37]

The literature also shows that people judge one another more harshly when they seek out algorithmic vs. human advice. Researchers at the University of Missouri designed an experiment in which participants read vignettes describing physicians who either did not seek advice, sought advice from a human expert, or sought advice from a computerized decision aid. Post-vignette surveys indicated that participants rated physicians who sought expert advice as significantly more positive than those who sought advice from a computerized aid or did not seek any advice at all. Thus, not only do people discriminate against algorithms in their own decisionmaking, but this study also showed that external observers are more likely to think negatively about those who rely on algorithms in their own decisions. All of this leads to an overall decreased willingness to use algorithms or computerized decision aids, and hints that algorithm aversion may become more acute as the work setting becomes increasingly complex.[38]

Conversely, other evidence indicates that some people tend to prefer complex algorithmic suggestions over human advice. In a set of experiments designed by a team of U.S. management psychology scholars, in which participants were asked to make decisions based on recommendations from "black box" algorithms alongside human recommendations, the participants exhibited a strong preference for the algorithms. Task domains included forecasting and making judgments about a visual stimulus. The results of these findings—an irrational preference for algorithmic over human advice—has been coined "algorithm appreciation."[39] Such preferences are not widespread and appear to often depend on a third variable resulting in algorithmic preference as a heuristic. For example, a German experiment's participants, who stated their belief that AI generally had greater intelligence than humans, were more likely to adopt algorithmic advice than those who believed humans were more intelligent.[40] In another set of experiments by a U.S. psychologist who specializes in big data, the participants preferred algorithmic advice when the decisionmaking domain required an objective vs. subjective judgment.[41]

This monograph uses the term algorithm reactance as a superordinate term to refer to phenomena generally related to both algorithm aversion and algorithm appreciation. However, the superordinate term is also

defined more broadly than algorithm aversion and appreciation *per se*, and refers generally to the entire class of attraction or repulsion effects associated with introduction of an AI agent into a decisionmaking setting.

Domain Expertise

The current research investigates the tangible drivers of AI algorithm reactance (RQ1), as well as ways to improve overall human AI interactions and joint decisionmaking (RQ2) in an IC setting. One relevant area in which the IC differs from other decisionmaking environments is in the type and specificity of employee domain expertise. Most research into joint human-AI decisionmaking relies on survey data from the public (e.g., Amazon MTurk or Prolific online panelists) or university students completing course lab requirements. While such respondents typically serve as a reasonable population from which to draw a sample, it is important to recognize that they are typically generalists and almost never experts in the domain in which they are being evaluated.

Research in the expertise literature has shown that this may be a significant shortcoming. Notably, previous work has shown that not only do experts process information differently, but they also respond differently than nonexperts. For example, research based on in-depth interviews and participant observation has shown that participants, as their expertise levels increase, more often make choices based on recognition rather than analysis.[42] This finding is further supported by empirical research evaluating the decisionmaking patterns of grandmaster chess players. Not only are expert players able to more easily retain large amounts of information in their working memories—indeed, many such players have been known to play blindfolded so their perceptual access to the chess positions is limited to working memory—but they were able to do so much more quickly and accurately than nonexperts, suggesting reduced processing time.[43, 44]

Furthermore, relative to nonexperts, experts' knowledge and reasoning tended to be more crystallized as a function of repetitive exposure and ongoing consolidation. For example, research into the phenomenon of consolidation—or the way in which people convert short-term memories and skills into long-term ones—shows that, after only a few minutes of hand-eye coordination tasks, participants demonstrated increased precision when completing similar tasks. After returning the next day, the same participants performed even better than they had at the end of the previous day. This suggests a cumulative effect in which long-term repetitive exposure to a set of tasks results in both short- and long-term gains in which greater expertise in a given domain depends on experience and memory. Thus, when solving a problem, people who consider themselves experts rely more on their own cognitive processes than on external cues. While nonexperts tend to reason their way through a problem, experts tend to recognize their way through one.

Note that IC intelligence analysts possess both formal education, such as academic degrees and certificates in which general knowledge and problem-solving skills are cultivated, and in-depth training specifically related to their chosen intelligence discipline. Furthermore, they are required to periodically complete additional training related to their disciplines, and (ideally) spend a majority of their days working in their assigned analytic domain. Thus, after only a few years, IC intelligence analysts possess both the formal (i.e., education, training, and on-the-job mentorship) as well as informal experience (i.e., thousands of hours of exposure to their analytic domain) necessary to invoke the cognitive processing patterns typified in both

crystallized and recognition-based decisionmaking characteristic of experts. Therefore, they may respond quite differently than nonexperts to external suggestions such as those from an AI agent.

Environmental Factor: User Interface Settings

The judgment and decisionmaking literature is replete with examples in which seemingly innocuous environmental factors influence human decisionmaking. For example, in both laboratory and field experiments people have been shown to violate transitivity, or the premise that people should prefer the same options irrespective of irrelevant contextual circumstances. Such effects have been shown in settings involving gambles with different outcomes, hypothetical admissions decisions of college applicants, and even real-life brand preferences.[45, 46, 47] Interestingly, despite the apparent economic value of preference transitivity, such violating behavior is not only commonplace but even extends to field observations of the animal kingdom.[48] Thus, the influence of contextual or environmental factors on cognition in general, and decisionmaking in particular, can be significant.

These findings may come as no large surprise to most people. Few consider themselves immune from seemingly irrelevant factors that can influence their decisionmaking. Parents are familiar with how a child's bad mood in the morning can influence their driving behavior while commuting to work, and most people are familiar with how skipping a meal can influence a decision one might make later in the day. Undoubtedly, most people are aware of their "triggers" and try to guard against how these environmental factors may influence their behaviors—and most people are generally successful.

However, some environmental factors may not be as obvious and, therefore, may be more likely to influence decisionmaking. User interface settings on a computer or decision support system is an example of one such area. Research considering this area is not new, but awareness of the insidiousness of its influences may not yet be widespread. For example, human factors research into the design and functionality of user interface settings for the Xerox 8010 personal computer resulted in design choices regarding the optimal number of buttons on the pointing device, the meanings for the buttons in the text-selection process, and the best icons to show users on the screen.[49] Nevertheless, instances abound in which failures to account for user interface design and other seemingly innocuous environmental factors result in inadvertent outcomes. In one such example, a poorly designed user interface resulted in a physician over-ordering medication from the Computerized Physician Order Entry system resulting in a medication overdose.[50]

In the IC, algorithms and software are commonly designed in-house. This is a function of the difficulty in securing approval for additions to software whitelists and of the need for highly idiosyncratic real-world applications, which may be classified. The inability to rely on widely available, well-tested, and documented commercial software means IC managers and developers shoulder the burden not only for developing in-house algorithms and software but also for minimizing the influence of extraneous environmental factors on end users as well.

Although most users probably can adapt to long-term influences generated by design missteps involving user interface settings, some environments and types of work may preclude the opportunity for people to successfully adapt. Over time, people have demonstrated successful adaptation to a wide variety of phenomena including shock in an experimental setting, loss of vision, lotteries, and perceptual judgments

such as customer satisfaction and happiness.[51, 52, 53, 54] In addition, people who work on the same computer system each day can individualize and standardize their user interface settings experience. In other environments, however, such as operations centers, employees routinely switch computers depending on factors including shift assignments, tasks, and mission requirements. These employees may work with systems with different user interface settings under time constraints that do not allow them to adapt.

Furthermore, in these less adaptive settings, employees are often on temporary assignment and vary widely in both their expertise and previous experiences. Recall that experts have crystallized knowledge and "recognize" rather than "reason" their way through decisionmaking, whereas nonexperts behave in the opposite manner. In a setting where both experts and nonexperts encounter user interface settings that are either congruent or incongruent from their previous experience, experts probably will continue to make decisions consistent with their established decisionmaking approach. That is, experts will be unaffected by any differences that changes to the user interface settings may induce. On the other hand, nonexperts probably will be influenced by the differences in user interface settings, as a function of their increased likelihood to "reason" their way through the decisionmaking process. Therefore,

Hypothesis 1 (H1): Experts will be less susceptible than nonexperts to deviations in *User Interface Settings*.

Human-in-the-Loop Hybrid Systems

As noted above in the background section, decreasing technology costs have resulted in significant increases in the amount of data the IC collects and then must sift through and evaluate. This torrent of information has widened the gap between mission requirements and human analytic capacity. AI agents offer the ability to process large volumes of data in environments in which the growth in data collection rate outpaces the growth in human analytic capability.

Returning to an earlier IC work setting example, one way in which AI is employed in the IC is to augment human analytic efforts in time dominant work environments such as operations centers or watch floors. Here, the goal is to generate a "first look" assessment of data before it is evaluated by analysts. This "first look" approach allows for timely triaging of the vast amount of data that is collected by various platforms, followed by further analysis as necessary.

Project Maven is one such example. The genesis of this Department of Defense program was to assist in processing the vast amount of full-motion video (FMV) of Islamic State militants in Iraq and Syria collected by various intelligence, surveillance, and reconnaissance (ISR) platforms. Following collection, an algorithm analyzes the data to detect objects or events modeled from a predetermined set. When the algorithm detects a target object, it flags the record, provides an assessment, and refers it for further human analysis and decisionmaking. The goal for the program is to allow AI agents and humans to work together "symbiotically to increase the ability of weapon systems to detect objects."[55]

There is widespread appeal in such human-in-the-loop hybrid systems, which employ the concept of ensuring that AI agents are not independent decisionmakers but must also include a human in the decisionmaking

process. Although AI-powered weapon systems are a necessary component of the U.S. defense strategy's goal to maintain a competitive edge with near-peer competitors, there has long been a cultural aversion to autonomous weapon systems. Such systems conjure images of killer robots and Skynet—the homicidal AI system from the Terminator movies.[56] And this concern is not entirely unfounded. Unchecked, AI-based systems have been shown to perpetuate biases resulting in racist, sexist, or even classist employment-related decisions.[57, 58, 59] AI-powered systems have also optimized protection of human life at the expense of non-human life,[60] but one can imagine a world in which they optimize some other variable at the expense of human life and dignity. Simply adding a human-in-the-loop, however, has been shown to improve some of these outcomes by increasing human salience toward the bias and even to improve AI processing time and accuracy by allowing human heuristics to effectively reduce the size of the decision search space.[61, 62]

Nevertheless, the question remains: If, as hypothesized above (H1), the introduction of an AI agent can influence human decisionmaking differently depending on individual differences and environmental factors, and if the addition of humans into the AI decisionmaking process influences the AI agent's decisionmaking, how do these two elements interact? Furthermore, what are some of the ways in which joint human-AI decisionmaking can be improved?

Explainable AI (XAI)

A current effort underway to improve joint human-AI interaction is explainable AI (XAI). Broadly, in order to improve human adoption of human-AI collaboration, XAI seeks to reduce the complexity that humans perceive in the recommendations offered by AI agents. XAI does so by designing AI agents that provide recommendations to their human counterparts that are relatively more interpretable.[63] For example, a recommendation that is not explainable may simply include the recommendation or conclusion based on the underlying dataset. However, a recommendation that is explainable may also provide the reasoning or "show the work" behind the conclusion.

Current approaches to improving explainability include four principles, as laid out by the National Institute of Standards and Technology: Explanation, Meaningful, Explanation Accuracy, and Knowledge Limits. AI agent recommendations should provide accompanying evidence on which a decision was based (Explanation), the explanation should be understandable to end users (Meaningful), it should reflect the agent's decision processes (Explanation Accuracy), and the agent should provide recommendations only for situations for which it was designed (Knowledge Limits).[64]

Note that how one defines "explainable" may vary greatly depending on personal preference, mission, or the cultural environment. For example, to extend the example provided above, developers might provide a link near the output or results of an AI agent's recommendation that users can click to learn more about the processes contributing to the conclusion. Or developers might provide an extensive set of output that provides all of the background information users might need to understand the AI agent's decisionmaking processes.

As previously discussed, research has shown a significant amount of variance in how people behave when interacting with AI agents. Some people irrationally prefer to work with AI agents, whereas others eschew

their use. Environmental and psychological factors also may influence cognitive processes and thus human interactions with an AI agent and subsequent decisionmaking. Because the use of algorithms such as AI agents has been shown to generally improve overall decisionmaking, however, any gain in usage is preferred.[65] And following XAI procedures should foster this gain by mitigating the natural distrust of AI agents that some people have, since providing more interpretable reasoning with AI agent recommendations has been shown to increase reported trust in the algorithm and lead to increased usage.[66, 67]

Based on this reasoning, the following additional hypotheses can be offered: AI agent recommendations framed with a High Explainability recommendation will increase task engagement, humans will subsequently rely on the AI agent more (Actual AI Reliance), and this greater reliance will result in increased overall quality of joint human-AI decisionmaking. Thus, the hypotheses to be tested in the author's research study include:

H2: A High (Low) Explainability recommendation from an AI agent will result in increased (decreased) accuracy.

H3a: Participants will perceive they rely on the AI agent more (less) in the High (Low) Explainability condition.

H3b: Participants will actively rely on the AI agent more (less) in the High (Low) Explainability condition.

H4: A High (Low) Explainability recommendation from an AI agent will result in more (less) task engagement.

Finally, because XAI has been shown to improve human trust in AI agents, people may exhibit less aversion toward the AI agent after interacting with it. Therefore, a fifth hypothesis can be added to the research study:

H5: A High (Low) Explainability recommendation from an AI agent will result in less (more) dislike toward the AI agent.

Recent efforts to explore and implement XAI systems in the IC have met with relative success, suggesting support for some of these hypotheses. For example, in 2016 DARPA launched an XAI program with the goals of creating AI models with more interpretable recommendations and subsequently fostering improved trust in the human-AI interaction.[68] And, across industry, various companies are offering services designed to augment existing AI agent services with XAI.[69]

However, even the best efforts to improve explainability may not be sufficient to meet end user needs for more complex deep learning algorithms. As described above, the most complex AI algorithms adapt to the underlying data in much the same way that a human is able to adapt routines to fit changing circumstances. For this reason, such algorithms are incredibly powerful and are sought after by mission owners seeking to emulate human decisionmaking, and they are used in both government and industry applications such as computer vision (CV), natural language processing (NLP), and generative adversarial networks (GAN). These algorithms provide significant benefit in the areas of feature detection, distillation of textual data for processing and reporting, and identification and generation of deepfake pictures and video, respectively. However, their underlying processes are also very complex. They have been characterized as explanatory

"black boxes" because an explanation of how they have reconfigured to evaluate the underlying data may not be intuitively accessible to even those who designed the algorithm. Ironically, in such cases the benefits of more complex algorithms may be abandoned in favor of simpler ones. Additional tools may be needed to improve end user acceptance of AI systems.

The Power of Choice

The literature review further suggests that, in addition to *Explainability* under the XAI thesis, allowing end users a *Choice* in how they receive AI agent recommendations may also play an influential role in whether they accept AI agent recommendations. Because allowing *Choice* may also play a significant role in decision accuracy and task engagement, it may serve as an important tool in the algorithm developer's toolkit, in addition to *Explainability*.[70]

Choice has been shown to be an operant factor in a number of decisionmaking contexts. For example, a study to evaluate the relative benefits of allowing examinees to select their own test items from a bank of similar test problems revealed that, when participants were given a *choice* of test items, test validity was enhanced by reducing response variance. Notably, the test items were nearly identical and differences in the actually chosen test items were nominal. Furthermore, participants preferred the test items they selected and actually performed more accurately—subsequently receiving higher scores.[71]

In an anagram task, when children were allowed a choice in the type of anagrams they would tackle, they solved significantly more anagrams than did children who were not allowed a choice. Moreover, those given a choice demonstrated significantly more intrinsic motivation. Interestingly, cultural differences (another environmental factor) were found to play a significant interactive role in performance and engagement as well, suggesting the relative influence of *Choice* on performance and engagement may be both environmentally determined and a learned behavior.[72]

And in an adversarial bargaining setting in which participants were asked to reflect on either their own choice options or those of their competitors, negotiators primed with a choice mindset perceived greater room for negotiation and were more willing to persist in negotiation than those not primed with a choice mindset—outcomes that are generally considered positive in this decisionmaking domain. Notably, in this set of U.S.-China business management studies, participants' choice options were not actually manipulated, but rather perception of the salience of choice was manipulated through a choice mindset prime.[73]

These findings are consistent with popular belief as well. Tag lines such as "don't give others power over your life" and "follow your values" abound in the popular business literature.[74] Thus, not only does a significant body of research support the idea that allowing people a choice is beneficial, but choice may be an expectation.

This monograph posits that many of the same improvements in task engagement and performance found associated with other task domains (e.g., bargaining, anagram completion, and test taking) may also translate to human receipt of recommendations from an AI agent. In particular, this paper hypothesizes

that providing a choice of recommendation output format (i.e., High vs. Low Explainability) to select study participants will result in their increased task engagement relative to the engagement of those study participants who did not receive a choice. Therefore, the following hypotheses are added to those already introduced:

H6: Allowing participants a choice (no choice) in Explainability Level format will result in increased (decreased) task engagement.

Furthermore, consistent with the choice literature and the linkage between engagement and performance, people may perform better and exhibit less dislike for working with the AI agent when offered a choice vs. no choice:

H7: Allowing participants a choice (no choice) in Explainability Level format will result in increased (decreased) accuracy.

H8: Allowing participants a choice (no choice) in Explainability Level format will result in decreased (increased) dislike for working with an AI agent irrespective of Explainability level.

The next section outlines the methodological approach used to address the above hypotheses on human-AI teaming (see Table 1). In particular, two sets of experiments are outlined. The first set of experiments investigates how an environmental factor, *User Interface Settings*, affects the extent to which people of varying degrees of domain expertise are influenced by the AI agent's recommendation. The second set of experiments considers not just influence effects but also explores an approach to improve joint human-AI decisionmaking. In particular, the effect of *Choice* on *Explainability Level* is investigated.

Table 1: Consolidated List of Hypotheses for Two Research Studies on Human-AI Teams

Study Set 1	
H1	Experts will be less susceptible than nonexperts to deviations in *User Interface Setting*s.
Study Set 2	
H2	A High (Low) Explainability recommendation from an AI agent will result in increased (decreased) accuracy.
H3a	Participants will perceive they rely on the AI agent more (less) in the High (Low) Explainability condition.
H3b	Participants will actively rely on the AI agent more (less) in the High (Low) Explainability condition.
H4	A High (Low) Explainability recommendation from an AI agent will result in more (less) task engagement.
H5	A High (Low) Explainability recommendation from an AI agent will result in less (more) dislike toward the AI agent.
H6	Allowing participants a choice (no choice) in Explainability Level format will result in increased (decreased) task engagement.
H7	Allowing participants a choice (no choice) in Explainability Level format will result in increased (decreased) accuracy.
H8	Allowing participants a choice (no choice) in Explainability Level format will result in decreased (increased) dislike for working with an AI agent irrespective of Explainability level.

Research Methodology: Understanding Human-AI Reactance and Team Performance

Study Set 1: Human-AI Reactance

The first set of studies of human-AI interaction—focused on human-AI reactance, or human avoidance of (attraction to) AI agents—considers a psychological factor, *Self-Assessed Expertise* (i.e., human participants judge themselves as expert or nonexpert on the subject of interaction with the AI agent), and an environmental variable, *User Interface Settings* (i.e., the format used by the AI agent to present recommendations). The overall goal is to investigate how S*elf-Assessed Expertise* interacts with *User Interface Settings* to understand whether individual differences in a psychological factor may influence human receptivity to AI agent recommendations when differences in an environmental variable are introduced. To this end, this study set manipulates participant *Self-Assessed Expertise* and evaluates how two different levels of this variable interact with two different levels of *User Interface Settings* to influence human-AI decisionmaking.

To accomplish this, a decisionmaking task was designed to serve as an analog for intelligence analysts deciding whether to accept an intelligence product based on the recommendation of an AI agent. Specifically, participants were asked to play the role of film studio executives deciding whether to produce a film based on elements presented in a (fake) film poster. In this study set, the participants' goal was to make decisions to either "begin" or "cancel" film production consistent with what the majority of other respondents chose (i.e., produce film posters that were popularly produced by other respondents or cancel production on film posters not popularly produced by other respondents). Note that the overall task relied on a domain generally familiar to most participants (films), but required them to make a specific decision in a domain with which they probably were unfamiliar (film production). In fact, while most participants indicated that they had watched a lot of films ("I watch a lot of films," Likert Scale, 1-7, Disagree to Agree, $M = 5.42$, $SD = 1.70$), few participants reported any significant actual experience in film production ("I have worked in film production," Likert Scale, 1-7, Disagree to Agree, $M = 1.32$, $SD = 1.09$). This familiarity-unfamiliarity dichotomy was critical to the *Self-Assessed Expertise* manipulation and is further discussed below.

Following recruitment into the study, participants were provided introductions that included an overview of the task and incentive structure (see Appendix 1: Study Set 1, Overview of Task, Instructions, and Incentive

Structure). Participants were then provided a set of "tips" designed to help them understand which poster elements typically resulted in a film doing well postproduction (i.e., typically resulted in the film being popularly produced). Participants subsequently took a test that purportedly assessed their understanding of the "tips" (see Appendix 2: Study Set 1, Information Set/Tips), but also served as an attention check (see Appendix 3: Study Set 1, Attention Check), and a series of practice decision tasks purportedly constructed to prepare them for the actual decision tasks. In fact, the "tips" and attention check formed the core of the *Self-Assessed Expertise* manipulation, and the way in which the practice and actual decision tasks were presented formed the environmental variable, *User Interface Settings*. A detailed description of both manipulations is provided below.

In the practice and actual decision tasks, participants were shown a set of film posters, each accompanied by an AI agent recommendation suggesting the film would be well or poorly received and were then asked to decide whether to "begin" or "cancel" production (see Appendix 4: Study Set 1, Sample AI Agent Recommendations Following Tutorial). In the first experiment in this study set all recommendations were presented in terms of a positive valence (e.g., "recommend begin production"). However, to introduce a more realistic recommendation environment, a second experiment presented recommendations in terms of both positive and negative valences (e.g., "recommend begin production" as well as "recommend cancel production"). In both experiments, presentation order of the film posters was randomized to prevent order effects. Participant decisions were incentivized by providing them a fixed incentive for study participation ($2.00) and increasing or decreasing their earnings by $0.05 for each decision that was consistent (i.e., correct) or inconsistent (i.e., incorrect) with majority opinion. Majority opinion was established through a pretest. Thus, the experiments were incentive-compatible in that participant choices were consequential, and performance in a manner consistent (inconsistent) with majority opinion resulted in increased (decreased) earnings.

User Interface Settings Manipulation

The *User Interface Settings* variable was constructed by randomly assigning participants to receive AI agent recommendations that were either verbal (e.g., "*Artemis* suggests that it is likely this film would do well if you begin production") or numeric (e.g., "*Artemis* suggests a 75-percent probability this film would do well if you begin production"), and presenting them in a pattern that was either congruent (e.g., verbal–verbal or numeric–numeric) or incongruent (e.g., verbal–numeric or numeric–verbal) across the practice and actual decision tasks. In this way, participants received an AI agent recommendation pattern in the actual decision task that simulated a *User Interface Settings* pattern that was either consistent (i.e., congruent) or inconsistent (i.e., incongruent) with their experience in the practice decision task.

Self-Assessed Expertise Manipulation

As mentioned briefly above, the *Self-Assessed Expertise* variable was constructed by randomly assigning participants to receive different information sets or "tips" that were either relevant or irrelevant to a subsequent attention check (see Appendix 2: Study Set 1, Information Sets/Tips). Note that the "tips" were

designed not to influence participant decisions related to any one film poster. Recall that the task's topic domain was selected to be generally familiar to participants (e.g., films) but the task itself was selected to be unfamiliar to participants (e.g., production decisions). This dichotomy allowed the experimenter to exploit the availability heuristic (i.e., the human tendency to mistake the ease or fluency with which a topic can be recalled with other assessments; here, Self-Assessed Expertise Level) so the participants' performance in the comprehension test manipulated their relative level of *Self-Assessed Expertise*.[75]

In fact, after calibration in a pretest, participants who received "tips" that were relevant assessed themselves as having a higher level of *Self-Assessed Expertise* than those who received irrelevant tips, suggesting that the manipulation was successful. Participants assigned to be Experts not only reported relatively higher levels of *Self-Assessed Expertise* than those assigned to be Nonexperts, but their ratings in three Self-Assessed Expertise manipulation check questions also crossed the scale midpoint (see Appendix 5: Study Set 1, Self-Assessed Expertise), suggesting they perceived themselves as actual experts in the task.

Thus, all participants were randomly assigned into a 2 x 2 (*User Interface Settings*: Congruent/Incongruent and *Self-Assessed Expertise*: Expert/Nonexpert) between-subjects experimental design, in which all subjects were randomly assigned to balanced groups, and the *User Interface Settings* variable was counterbalanced to account for possible order effects in its subordinate factor (Presentation Mode). The primary dependent variable was *Number of Decisions Accepting AI Recommendations* in the film production choice task.

Note that this study set was designed to investigate factors that influence human decisionmaking when working as part of a human-AI team. A more thorough understanding of which factors may inadvertently influence individual-level analyst decisions is crucial for IC mission owners seeking to augment their existing human workforce with AI agents. Failure to account for these effects may lead to different decisionmaking outcomes than would otherwise have been made. Note, however, that this study set was designed so there were no "right" or "wrong" answers *per se*—participants were simply asked to provide responses generally consistent with popular opinion. Although some decision environments involve ambiguous decisions like these, many analytic environments have relatively clear "right" and "wrong" answers. Furthermore, this study set was artificial because the sampling plan was set up with professional respondents from the public participating in a contrived decisionmaking setting (e.g., film production). Although most research generally shows that such studies tend to produce valid and generalizable results, some research does show that laboratory and field studies can produce dramatically different results.[76, 77]

To broaden this exploration of human-AI reactance, a second study set explored ways to improve the quality of joint human-machine decisionmaking. What current approaches can improve human-machine cooperation? What can managers do to improve actual decision-outcome quality for human-machine decisionmaking teams?

Understanding how an AI agent's interactions might influence its human counterparts' accuracy is critically important for managers seeking to implement human-machine teaming in their mission spaces. The next study set addresses these issues in the context of XAI principles, which propose that the way AI agents arrive at their recommendations should be understandable to their human counterparts.[78] Thus, Study Set 2 will

manipulate *Explainability Level* (High vs. Low), *Choice* of *Explainability Level* (Choice vs. No Choice) and measure how well participants perform in a task (*Accuracy*), how much they rely on the AI agent (*Actual AI Reliance*), and their own level of awareness of their reliance on the AI agent (*Self-Assessed Reliance*).

Study Set 2: Human-AI Team Performance*

The second study set considers two different factors, discussed in the literature review, that may improve the overall performance of human-AI teaming: *Explainability Level* and *Choice*. The goal is to investigate how both *Explainability Level*, which is derived from the XAI thesis, and *Choice* impact decisionmaking processes and outcomes in a set of joint human-AI tasks. To this end, Study Set 2 includes two experiments that manipulate *Explainability Level* and *Choice* to assess (1) how these factors independently influence decisionmaking and (2) how they interact with one another. The two experiments share a common experimental design, although some deviations from this design will be noted in the discussion of Experiments 1 and 2 below.

Common Experimental Design

Following recruitment into the study, participants were given instructions on the task purpose and their scope of responsibilities in the study. Generally, participants were asked to review either an image or a text document and to count experimenter-designated attributes within the document. Tasks involving images and text were selected based on their similarity to actual tasks performed by IC analysts.

In the study instructions (see Appendix 7: Study Set 2, Experiment 1, Study Instructions), participants were told that during the task they would receive assistance from an AI agent using the "latest algorithms." Specifically, the agent would perform the same task and generate a preferred solution using six different algorithms. The algorithmic output was artificial, although this was not known to participants. Furthermore, participants were told that, although everyone would receive the AI agent's recommendation, the extent to which they used this recommendation in their own responses was a personal choice.

The experimenter then showed participants a sample of the user interface and allowed them to complete an example task. Participants were randomly assigned to receive a task orientation and example in which an AI agent recommendation was either in a Low Explainability format (i.e., the AI agent autonomously selected the preferred solution from the set of six algorithms; see Figure 3, Panel A) or in a High Explainability format (i.e., the AI agent provided output from the set of six algorithms and allowed the participant to identify which of the six was preferred; see Figure 3, Panel B).

Participants who received the AI recommendations in a High Explainability format also received a short orientation to the output and how to interpret it. Specifically, they were told they could identify the

* The author acknowledges the material support of Erik Hatfield (NGA) and Dain Thomsen (USAF) in both design of the experiment and collection of the data.

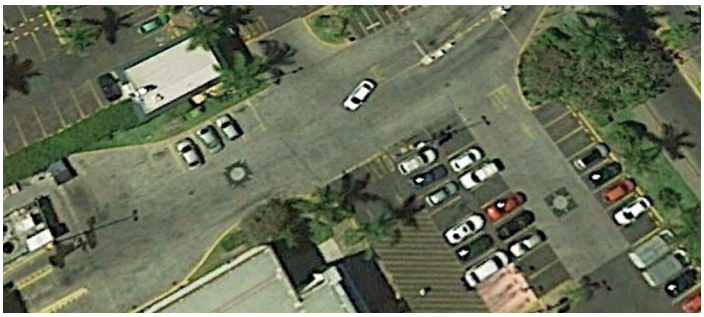
preferred solution by comparing the results of each algorithmic model 1-6, based on Akaike Information Criterion (AIC)—a mathematical method that evaluates how well a model fits the data from which it was generated—and Bayesian Information Criterion (BIC), which measures the trade-off between model fit and the model's complexity. Because the algorithm model with the lowest AIC and BIC values indicates best fit, that model generated the preferred solution (see Figure 3, Panel B, with the lowest AIC and BIC values highlighted in blue and the corresponding algorithmic model recommendation highlighted in green). Participants were then free to perform the task, accept the AI recommendation as they saw fit, and provide their answers. After this orientation, participants performed a series of actual tasks. At the conclusion of these tasks, participants completed a short survey assessing behavioral factors, and then were dismissed.

Accuracy was assessed by comparing participant responses to known solutions and noting the number of deviations. *Actual AI Reliance* was measured by manipulating the AI agent's output such that its recommendations were stochastically determined following a uniform distribution with a mean equal to the task's known solution. Thus, at an individual level, AI agent recommendations were unrelated to the "correct" solutions and, therefore, could not bias sample-level results; however, at a sample level, participants received AI agent recommendations that were statistically equivalent to the correct answer. This enabled the experimenter to simultaneously analyze *Accuracy* at a sample level by calculating

Figure 3. Examples of Explainability Levels

Panel A: Low Explainability

Below, is an example of the type of image you will see.
Remember, in this study we wish to know how many four-door sedans you can find.

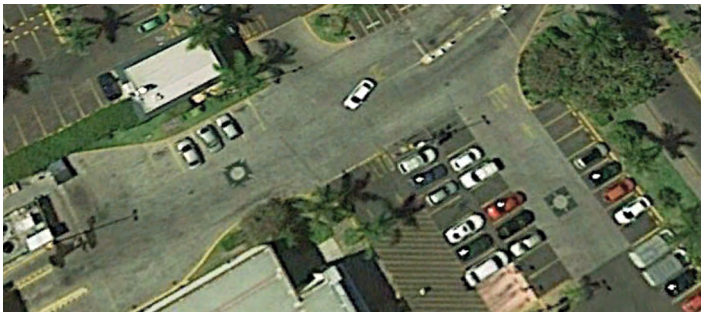


AI Assessment (Number of Four-Door Sedans): 4

Your Assessment (Number of Four-Door Sedans): ____

Panel B: High Explainability

Below, is an example of the type of image you will see.
Remember, in this study we wish to know how many four-door sedans you can find.



AI Assessment (Number of Four-Door Sedans):

	Determinants of the Logarithm of Vehicle Count					
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
# of Vehicles	0.32 (0.15)	0.74 (0.26)	0.13 (0.03)	0.67 (0.22)	0.87 (0.31)	0.76 (0.27)
Length	0.55 (0.35)	0.78 (0.27)	0.87 (0.31)	0.33 (0.11)	0.81 (0.22)	0.97 (0.34)
Shadow Incidence	0.43 (0.12)	0.91 (0.37)	0.83 (0.19)	0.66 (0.42)	0.79 (0.15)	0.65 (0.38)
Angle (X) Incidence	0.87 (0.44)	0.38 (0.28)	0.86 (0.37)	0.28 (0.01)	0.18 (0.04)	0.07 (0.01)
Angle (Y) Incidence	0.75 (0.25)	0.01 (0.00)	0.15 (0.03)	0.43 (0.13)	0.77 (0.32)	0.31 (0.15)
Angle (Z)	0.23 (0.11)	0.31 (0.15)	0.33 (0.31)	0.45 (0.01)	0.65 (0.15)	0.83 (0.11)
Size	0.08 (0.01)	0.11 (0.05)	0.44 (0.12)	0.23 (0.11)	0.17 (0.03)	0.19 (0.13)
Shape	2L	2L	2L	2L	2L	2L
(-2 Res Log Likelihood)	3853	3824	4351	3421	3555	3422
AIC (Smaller is Better)	3543	**3122**	3129	3475	3659	3239
AICC (Smaller is Better)	3846	3827	3449	3495	3963	3227
BIC (Smaller is Better)	3727	**3112**	3447	3118	3927	3857
Null Model						
Likelihood	0.000174	0.000182	0.000144	0.000973	0.000835	0.000145
Ratio Test						
*SE in Parentheses						
Estimates:	Model 1: 7	Model 2: **4**	Model 3: 26	Model 4: 8	Model 5: 6	Model 6: 24

Your Assessment (Number of Four-Door Sedans): ____

Source: "Parking lot of Restaurante Mar y Tierra Veleiros, Jalisco, Mexico," Google Maps, accessed on March 14, 2021, https://www.google.com/maps/@20.709417,-103.41092,45m/data=!3m1!1e3.

how far participant solutions deviated from the known solutions, and *Actual AI Reliance* at an individual level by calculating how far participant solutions deviated from the AI agent-provided recommendation.† For both *Accuracy* and *Actual AI Reliance*, fewer deviations (i.e., smaller numbers) represented increased accuracy or reliance—with zero being a perfect score. *Self-Assessed AI* Reliance was measured in the post-task survey using four Likert-scale survey items (see Appendix 8: Study Set 2, Self-Assessed AI Reliance).

Experiment 1: Explainability Level

Experiment 1's objective was to understand the extent to which AI agent-based recommendation formats (i.e., Low Explainability vs. High Explainability) might influence the quality of participant decisionmaking. This experiment employed an imagery-based task in which *Actual Correctness* served as a proxy measure for the quality of participant decisions. Specifically, participants were asked to count the number of four-door sedans in a Google Earth satellite image (see Figure 3 for an example). The author hypothesized that participants who received an AI agent recommendation in a High Explainability format would be more engaged in the task (H4) and, therefore, more likely to rely on the AI agent's recommendations (H3a-b), so that they would subsequently be more accurate than those who received an AI recommendation in a simple format (H2: see Table 1 for hypotheses list).

The task—counting the number of four-door sedans in a series of four satellite images—was designed to be difficult. The selected images were blurry, and they contained vehicles that looked like four-door sedans but may have been two-door sedans or hatchbacks. All images were from non-U.S. locations. To ensure participants did not mistake similar type vehicles as four-door sedans, participants were shown a set of similar vehicles contained within the image, clarifying that the target type was a four-door sedan. Post-task measures revealed that participants generally rated the task as moderately difficult: $M_{Complexity} = 4.28$, $SD_{Complexity} = 2.04$, r = 0.76. Prior to study execution, "correct" solutions were determined through interrater agreement: two geospatial intelligence (GEOINT) analysts reviewed each image used in the study and counted the number of sedans. Raters achieved 92.21 percent agreement and the remaining differences were resolved through discussion.

The overall study was a single-factor, mixed design in which AI Recommendation Complexity was between-subjects. The dependent variables *Accuracy, Actual AI Reliance,* and *Time* were measured within-subjects on multiple occasions, and *Self-Assessed AI Reliance* and *Dislike* were measured once.

Experiment 2: Choice of Explainability Level

Experiment 1 was designed to investigate the influence of the XAI thesis on task engagement, AI agent reliance (both perceived and actual), and subsequent performance accuracy in a GEOINT task. As noted in

† A 500-cycle bootstrapped simulation of 5,000 records was used to validate that this approach would generate AI agent-based recommendations that were uncorrelated (0.001), nearly completely random (0.95), and, therefore, orthogonal or unrelated to the known solutions, r(5000) = 0.001, p = 0.95.

the literature review, some situations may not allow greater explainability without also increasing complexity—an outcome that may increase both participant dislike toward the AI agent and subsequent nonusage. Thus, the goal for Experiment 2 was to explore an additional, relatively easy-to-implement factor that might influence AI agent receptivity—allowing participants *Choice* in their *Explainability Level*.

To add *Choice* as a factor, Experiment 2's instructions were redesigned so that all participants received an orientation to both the simple and complex recommendation formats, but only some participants were allowed to choose the type of recommendation format they would receive. All participants. therefore, were aware of the possibility of two different *Explainability Levels* (High and Low). Participants were then randomly assigned to have either a Choice or No Choice as to whether they would receive the recommendation in Low or High *Explainability Levels*. Consistent with Experiment 1, the second experiment captured (either during or in a post-experiment survey) the following dependent variables: *Accuracy*, *Actual AI Reliance*, *Perceived AI Reliance*, *Dislike*, and *Task Engagement*.

Experiment 2 also used a somewhat different task. Whereas Experiment 1 was an imagery-based task in which participants were asked to count four-door sedans, Experiment 2 was a text-based task in which participants were asked to read a passage and count the number of errors. Consistent with the sampling approach used in Experiment 1, the author recruited a sample of IC business analysts and editors who were familiar with the task domain.

Thus, Experiment 2 was a 2 x 2 (*Choice*: Choice or No Choice and *Explainability Level*: High vs. Low) mixed design, in which *Choice* and *Explainability Level* were between subjects. *Accuracy*, *Actual AI Reliance*, and *Time* were measured within subjects on four occasions, and *Perceived Reliance* and *Dislike* were measured once at the conclusion of the study.

Findings: Choice of Algorithm Output Complexity Improves Overall Human-AI Team Compatibility and Performance

AI Recommendations Influence Nonexpert Decisionmaking More When the User Interface Is Unfamiliar (Study Set 1)

Consistent AI Agent Recommendations (Experiment 1)

The goal for Study Set 1, as noted in the previous section, was to investigate how participants of differing levels of expertise (*Self-Assessed Expertise*) responded to the introduction of an AI agent making recommendations using either similar or different formats across occasions (*User Interface Settings*). To accomplish this, 103 online panelists were recruited from Amazon's crowdsourcing marketplace, Mechanical Turk, and randomly assigned to the two levels of *Self-Assessed Expertise* ($n_{Experts}$ = 51; $n_{Non\text{-}experts}$ = 52) and to the two levels of *User Interface Settings* ($n_{Congruent}$ = 57; $n_{Incongruent}$ = 46). In this first experiment, for simplicity, all participants received AI agent recommendations that were positively valenced (i.e., "Begin Production") rather than a mix of positively valenced and negatively valenced (i.e., "Cancel Production") recommendations. A summary of Study Set 1 results relative to the hypothesis is offered below and is discussed in depth throughout the rest of this section.

Table 2: Summary of Study Set 1 Findings

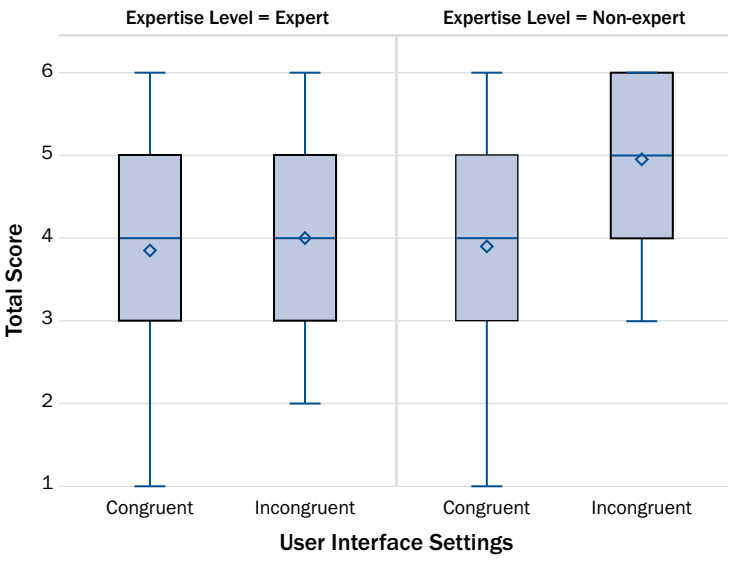
H1	Experts will be less susceptible than nonexperts to deviations in *User Interface Settings*.	**Supported**

Manipulation checks on the effectiveness of *Self-assessed Expertise* showed that the manipulations were successful. A set of three Likert-scale items designed to measure self-assessed expertise revealed that participants assigned to the Expert condition (M_{Expert} = 5.14, SD_{Expert} = 0.08) rated themselves significantly more expert

in making production decisions for films than those assigned to the Nonexpert condition (M_{Expert} = 1.60, SD_{Expert} = 0.07), F(1, 101) = 1116.59, p < 0.001 (see Appendix 5: Study Set 1, Self-Assessed Expertise). Importantly, ratings by participants assigned as Experts crossed the scale midpoint, indicating they assessed their level of expertise in the target domain in actual vs. merely relative terms. Measures of participant perceptions of *User Interface Settings* were not captured as these were environmental variables.

Before interpreting the results, the data were analyzed using a series of generalized linear models allowing for longitudinal binary responses (i.e., "Begin Production" or "Cancel Production"). Model estimation relied on maximum likelihood within SAS GLIMMIX (i.e., a statistical procedure that predicts population parameter values by quantifying the joint probability for predicting a given sample of data) to assess and select final models with the best distributional and variance-covariance matrix fits. The selected model predicted the number of positive production decisions, relying on a binary distribution with a logit link function (i.e., to keep the proportion of production decisions between 0 and 1), and the model also allowed random intercept variance, separate residual variances per occasion, and unstructured residual correlations for actual task outcomes. To identify the simplest, most powerful statistical analysis for modeling the data, the author evaluated several models to account for patterns of nonnormality of data as well as how differences in data variance across measurement occasions might exhibit time-based dependence because of participants' previous choices. A Poisson distribution, allowing for unstructured variance across measurement occasions, had the best fit. Note, outcome results are typically provided in "logits," or log-odds units, which are not further discussed here (see Chapter 3 of Craig Enders' *Applied Missing Data Analysis* for an excellent overview); however, for convenience the author has converted the results back to percentages.[79]

Figure 4. Number of Decisions Accepting AI Recommendation, Study Set 1, Experiment 1



An Analysis of Variance (ANOVA) test, which determines statistical differences between the means of independent groups, was applied to the participant production decisions. This revealed a significant interaction between *Self-Assessed Expertise Level* and *User Interface Settings*. As can be seen in Figure 4, Nonexperts were significantly more likely than Experts to agree with the recommendations from the AI agent when *User Interface Settings* were Incongruent rather than Congruent between the practice and actual tasks, F(1, 515) = 4.01, p = 0.07 (see also Appendix 10: Study Set 1, Experiment 1, ANOVA Results (Agreements)). This is contrary to expectation because one would normally expect any incongruency between the practice

and actual task to appear to be an error on the part of the algorithm, and previous research has shown that perceived errors reduce trust and subsequent use.[80]

One possibility for these anomalous ANOVA test results may be the outcome of the AI agent recommendations being presented in terms of only a positive valence (i.e., "this film would do well"). This was an intentional design choice to reduce response variance and improve the likelihood of detecting an effect. In real life, however, people are likely to receive AI agent recommendations with both positive and negative valences (i.e., "this film would do well" or "this film would not do well"). The consistent pattern of positively valenced recommendations may have encouraged participants who felt less confident in their decisions (e.g., Nonexperts whose confidence may have been manipulated by the *Self-Assessed Expertise* manipulation) to also be more likely to "Begin Production" every time.

If the observed pattern of results was a function of the artificiality of recommendation valence, one would expect a balanced presentation of recommendations to attenuate the results pattern. Furthermore, if the pattern of results was a function of erosion of participant confidence because of the *Self-Assessed Expertise* manipulation, then participants assigned as Nonexperts should demonstrate reduced confidence consistent with the observed pattern. Both of these propositions were tested in Study Set 1's second experiment.

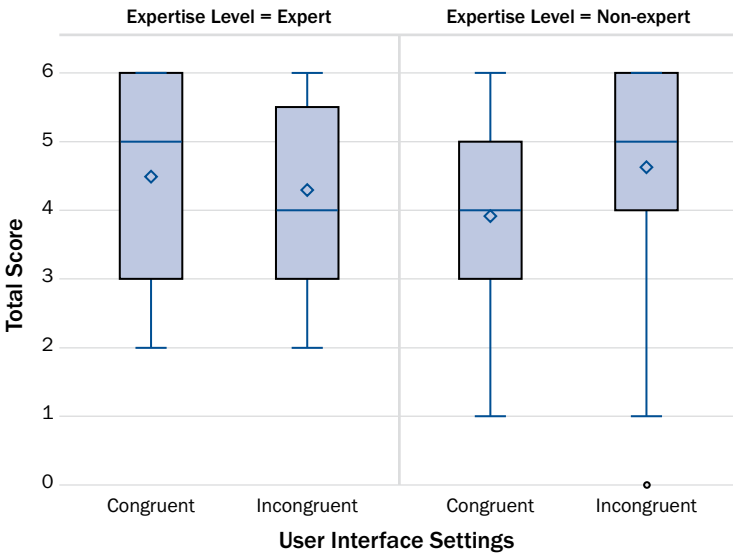
Inconsistent AI Agent Recommendations (Experiment 2)

The goal of Experiment 2 was to replicate the results from Experiment 1, while extending the experimental design to be more realistic and conducting a mediation analysis to investigate the roles that reliance and confidence may have on the observed effect. In Experiment 1, as reported above, Nonexperts who received AI agent recommendations in an Incongruent format between the practice and actual tasks were significantly more likely to accept AI agent recommendations. This pattern of results was surprising, and the extent to which this may have been influenced by participant reliance on the AI agent, self-confidence levels, or the realism of the decisionmaking environment was unclear.

To evaluate these possible influence factors, 157 online panelists were recruited from Amazon's crowdsourcing marketplace, Mechanical Turk, and randomly assigned to the two levels of *Self-Assessed Expertise* ($n_{Experts}$ = 81; $n_{Non\text{-}experts}$ = 76) and to the two levels of *User Interface Settings* ($n_{Congruent}$ = 73; $n_{Incongruent}$ = 84). The experimental design was modified to allow the AI agent to provide both positively and negatively valenced recommendations, and the sequence in which these recommendations were presented was randomized to prevent order effects. Furthermore, measures were included to assess the degree to which participants actively relied on the AI agent recommendation (see Appendix 11: Study Set 1, Experiment 2, Reliance Measures and Mediation Results) and on their own self-confidence (see Appendix 12: Study Set 1, Experiment 2, Self-Confidence Measures) in making their decisions.

As in Experiment 1, before interpreting the results, the data were analyzed using a series of generalized linear models allowing for longitudinal binary responses (i.e., "Begin Production" or "Cancel Production"). The same model (Poisson-distributed, unstructured variance across measurement occasions) remained the

Figure 5. Number of Decisions Accepting AI Recommendation, Study Set 1, Experiment 2



best-fitting predictor for the number of positive production decisions.

ANOVA results of this second experiment showed the observed pattern of participant decisions generally conformed to those from Experiment 1 (see Figure 5). As expected, introduction of both positively and negatively valenced AI agent recommendations attenuated some of the observed differences in the results; however, Nonexperts continued to accept AI agent recommendations significantly more often when receiving *User Interface Settings* that were incongruent across the practice and actual tasks (see Appendix 13: Study Set 1, Experiment 2, ANOVA Results (Agreements)).

A mediated moderation analysis was performed to investigate the possibility that the observed pattern of results was a function of participants actively relying on the AI agent's recommendation (i.e., reliance), as well as whether the Self-Assessed Expertise manipulation may have influenced participant self-confidence (i.e., confidence).[81] Results indicated that the *Self-Assessed Expertise* x *User Interface Settings* interaction term was a significant predictor of reliance, B = 1.13, SE = 0.17, $p < 0.001$, and that reliance further predicted the number of AI recommendations accepted by participants, B = 0.09, SE = 0.03, $p < 0.001$. After including reliance in the mediated moderation model, the *Self-Assessed Expertise* x *User Interface Settings* interaction term was no longer a significant predictor of the number of AI recommendations accepted by participants, B = 0.10, SE = 0.06, $p = 0.11$, suggesting reliance fully mediated the observed pattern of results (see Appendix 11: Study Set 1, Experiment 2, Reliance Measures and Mediation Results). In contrast, when confidence was included in the model, the *Self-Assessed Expertise* x *User Interface Settings* interaction term remained significant but confidence was not significant, suggesting the latter was not a mediator. In other words, Nonexperts were motivated by reliance, not by confidence, and their reliance on the AI agent recommendation was a conscious choice, not influenced by the experimental setup.

Human-AI Team Performance (Study Set 2)

Explainability Level (Experiment 1)

Study Set 2, as described in the Research Methodology section, was designed to explore ways to improve the overall performance of human-AI teaming, with a focus on the factors of *Explainability Level* and *Choice*. The goal of Experiment 1 in this study set was to investigate the influence of the Explainable AI

(XAI) thesis on participant task engagement, AI agent reliance (both actual and perceived), and subsequent performance accuracy when receiving AI agent recommendations in either a Low or High Explainability format. A summary of Study Set 2 results related to the hypotheses is offered below. Comprehensive analyses of the results are further discussed throughout this section.

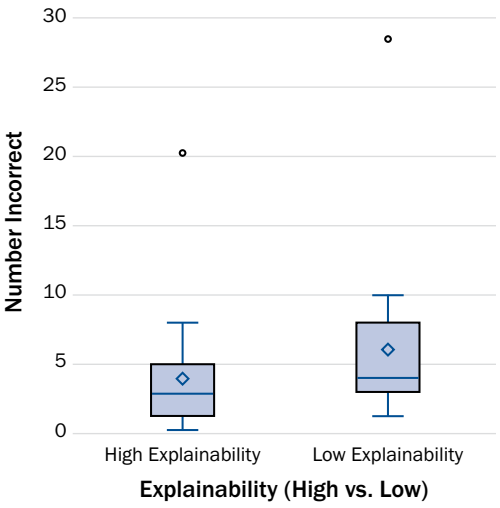
Table 3: Summary of Study Set 2 Findings

Study Set 2 Results		
H2	A High (Low) Explainability recommendation from an AI agent will result in increased (decreased) accuracy.	**Supported**
H3a	Participants will perceive they rely on the AI agent more (less) in the High (Low) Explainability condition.	**Not supported: opposite pattern**
H3b	Participants will actively rely on the AI agent more (less) in the High (Low) Explainability condition.	**Not supported**
H4	A High (Low) Explainability recommendation from an AI agent will result in more (less) task engagement.	**Not supported**
H5	A High (Low) Explainability recommendation from an AI agent will result in less (more) dislike toward the AI agent.	**Supported**
H6	Allowing participants a choice (no choice) in Explainability Level format will result in increased (decreased) task engagement.	**Supported**
H7	Allowing participants a choice (no choice) in Explainability Level format will result in increased (decreased) accuracy.	**Supported**
H8	Allowing participants a choice (no choice) in Explainability Level format will result in decreased (increased) disliking irrespective of Explainability level.	**Partially supported**

To this end, 75 analysts with a GEOINT background were recruited to participate in an experiment in which they were asked to view a series of four satellite images and provide a count of the number of four-door sedans seen in these images. Participants were randomly assigned to receive AI recommendations in either a Low Explainability (n = 37) or High Explainability (n = 38) format, and a total of 47 participants completed the full study. Attrition was not significantly different between the assignment conditions, χ^2 (0.75, n = 75) = 0.39. On average, participants were 37.77 years old (M_{Age} = 37.77, SD_{Age} = 10.61), held 8.16 years of imagery analyst experience ($M_{Experience}$ = 8.16, $SD_{Experience}$ = 7.28), self-reported as having relatively greater expertise on a seven-point, Likert scale item (Disagree to Agree, 1-7) measuring relative experience performing in GEOINT ("I am an experienced imagery/GEOINT analyst," $M_{SelfExpertise}$ = 5.36, $SD_{SelfExpertise}$ = 1.79), and self-reported as 63.83 percent male and 36.17 percent female. Checks to measure the *Explainability Level* variable's effectiveness showed it successfully manipulated participants' self-evaluations of expertise. Furthermore, participants did not find the actual task more (p = 0.84) or less (p = 0.35) difficult across Explainability conditions.

Before interpreting the *Accuracy* and *Actual AI Reliance* results, analytical power was maximized by optimizing model fit. The preferred models for *Accuracy* and *Actual AI Reliance* were Poisson-distributed, and *Time* and *Self-Assessed AI reliance* were Gamma-distributed.

Figure 6. *Accuracy* Results, Study Set 2, Experiment 1

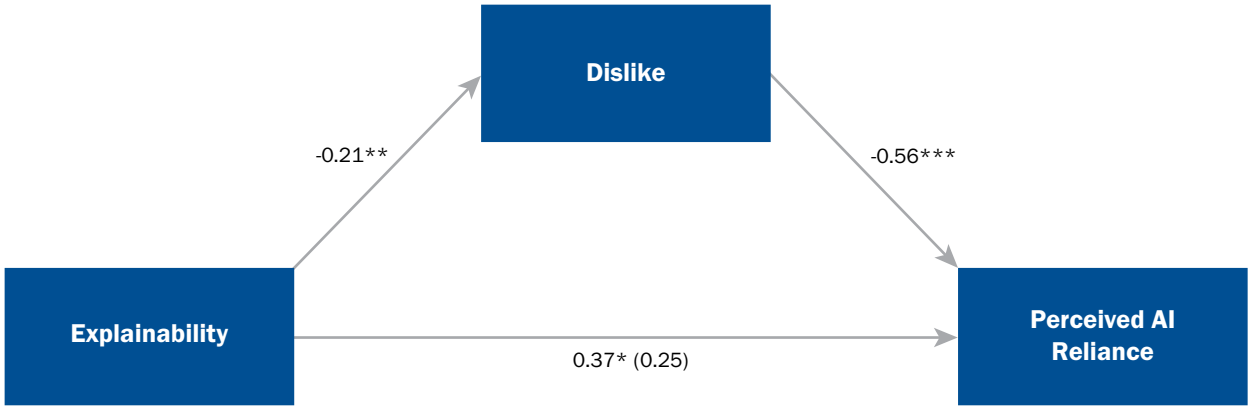


As expected, results for *Accuracy* (see Appendix 14: Study Set 2, Experiment 1, ANOVA Results (Accuracy)) revealed that the *Explainability Level* variable exercised a significant effect on the participants' decisionmaking acumen. Those assigned to the High Explainability condition were significantly more accurate ($M_{HighExplainability}$ = 3.94, $SD_{HighExplainability}$ = 0.58) than those assigned to the Low Explainability condition ($M_{LowExplainability}$ = 6.06, $SD_{LowExplainability}$ = 0.68), F(1, 45) = 5.37, p = 0.03 (see Figure 6).

Interestingly, participants assigned to the Low Explainability condition perceived themselves as relying more on the AI agent's recommendation ($M_{LowExplainability}$ = 3.72, $SD_{LowExplainability}$ = 0.13) than those assigned to the High Explainability condition ($M_{HighExplainability}$ = 3.35, $SD_{HighExplainability}$ = 0.14), F(1, 141) = 3.58, p = 0.06. Despite this perception, however, tests of *Actual AI Reliance* revealed that participants in the Low Explainability condition did not *actually* rely more on the AI agent's recommendations, F(1, 45) = 1.32, p = 0.26 (see Appendix 15: Study Set 2, Experiment 1, Actual AI Reliance Compared to Self-Assessed AI Reliance).

A further examination revealed that the covariate *Dislike* (i.e., dislike toward the AI agent) mediated the relationship between *Explainability Level* and *Perceived AI Reliance*. As Figure 7 illustrates, both the standardized regression coefficients between *Dislike* and *Perceived AI Reliance* and between *Explainability Level* and *Dislike* were statistically significant. Interestingly, these results showed that consistent with the XAI hypothesis greater Explainability predicted a significant decrease in Dislike felt toward the AI agent. Decreased dislike subsequently resulted in significantly increased perceived AI reliance.

Figure 7. Mediation of *Explainability* on *Perceived AI Reliance* by *Dislike*



* p < 0.10, ** p < 0.05, *** p < 0.01

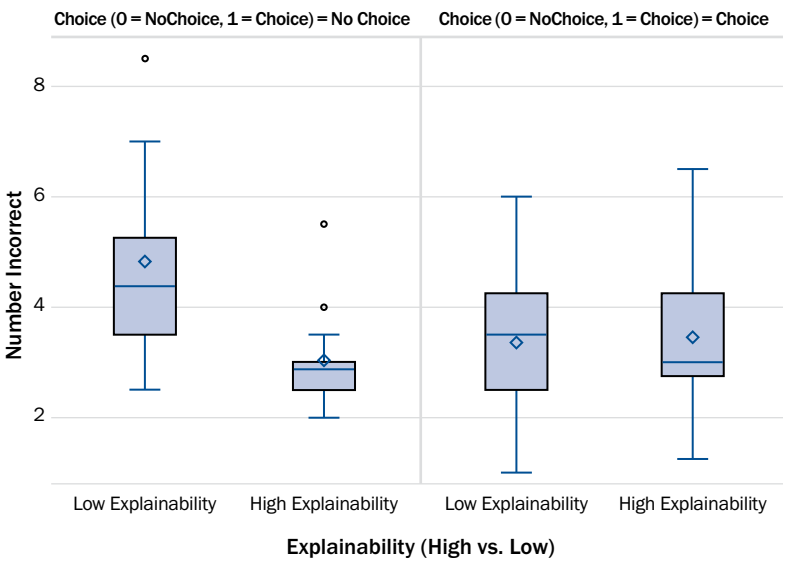
Choice of Explainability Level (Experiment 2)

The goal of Experiment 2 was to delve more deeply into the mixed results observed in Experiment 1, in which participants perceived greater (though not actual) reliance on Low Explainability AI recommendations, disliked High Explainability AI recommendations, and yet interestingly performed more accurately when receiving AI agent recommendations with High Explainability. Specifically, the objective was to investigate the influence of providing participants a *Choice* of *Explainability Level* on their perceived and actual reliance, task engagement, and subsequent accuracy.

To this purpose, 124 current and former business analysts and editors were recruited from within the National Geospatial-Intelligence Agency (NGA) and randomly assigned to the two levels of *Choice* ($n_{No\ Choice}$ = 62; n_{Choice} = 62); those in the No Choice condition were assigned to an *Explainability Level*, while those in the Choice condition were permitted to select their *Explainability Level*. Not surprisingly, a majority (68.1 percent) of those who were provided a choice preferred to receive their AI recommendations with a Low Explainability format, with the rest preferring to receive their recommendations with a High Explainability format.

As in Experiment 1, analytical power was maximized by reducing the number of estimated parameters and improving fit with respect to the response variable. Specifically, alternative variance-covariance models across occasions and with different response variable distributions (e.g., Poisson and Negative Binomial, again because of the count nature of the data) were examined. Nonnested alternate response variable distributions were fit by selecting the model with the generalized $\chi^2/d.f.$ closest to 1, and nested alternate variance-covariance models were fit using likelihood ratio tests. The preferred combination of response distribution and variance-covariance structure for both *Correctness* and *Actual AI Reliance* tended to be Poisson, distributed with separate residual variances per occasion, and Variance Components residual correlations for measurement periods 1-4. The preferred response variable distribution and variance-covariance structure for *Time* and *Self-Assessed AI reliance* tended to be Gamma, distributed also with separate residual variances per occasion, and Variance Components residual correlations for measurement periods 1-4.

Figure 8. *Accuracy* Results, Study Set 2, Experiment 2



Consistent with the results in Experiment 1, participants who received their recommendations with a High Explainability format made fewer errors ($M_{High\ Explainability}$ = 3.30, $SD_{High\ Explainability}$ = 0.13) than those who received recommendations with a

Low Explainability format ($M_{Low\ Explainability}$ = 3.90, $SD_{Low\ Explainability}$ = 0.12), F(1, 68) = 7.45, p = 0.008 (see Appendix 16: Study Set 2, Experiment 2, ANOVA Results (Accuracy)). Also consistent with Experiment 1, the *Explainability Level* variable did not significantly influence participants' actual reliance on the AI agent, F(1, 68) = 0.08, p = 0.78. Interestingly, *Explainability Level* did continue to predict participants' *Self-Assessed AI Reliance* to a significant degree, F(1, 67) = 11.24, p = 0.001, but in a direction opposed to the XAI argument that an AI agent providing recommendations that are relatively more interpretable will engender trust and, therefore, reliance. In contrast, participants who received AI recommendations in a Low Explainability format reported significantly greater reliance on the AI agent ($M_{Low\ Explainability}$ = 3.90, $SD_{Low\ Explainability}$ = 0.12) than those who received recommendations in a High Explainability Format ($M_{High\ Explainability}$ = 3.30, $SD_{High\ Explainability}$ = 0.13), F(1, 67) = 11.24, p = 0.001 (see Appendix 17: Study Set 2, Experiment 2, Actual AI Reliance Compared to Self-Assessed AI Reliance).

As seen in Experiment 1, mere *Choice* did not significantly influence participants' performance, F(1, 68) = 0.15, p = 0.15, or actual reliance on the AI recommendations, F(1, 68) = 2.17, p = 0.48 (see Appendix 14: Study Set 2, Experiment 1, ANOVA Results (Accuracy)). However, allowing participants a *Choice* in *Explainability Level,* as offered in Experiment 2, did influence their own assessment of how much they relied on the AI agent, F(1, 67) = 33.86, p < 0.001 (see Appendix 17: Study Set 2, Experiment 2, Actual AI Reliance Compared to Self-Assessed AI Reliance). Furthermore, provision of choice attenuated the mediation of *Explainability Level* and *Dislike* toward the AI agent that was observed in Experiment 1. Critically, Experiment 2's results also revealed a significant interaction between *Explainability Level* and *Choice:* providing *Choice* irrespective of whether a participant selected High Explainability or Low Explainability improved performance in the task to match that of the No-Choice/High Explainability condition, F(1, 68) = 9.62, p = 0.003. Finally, an analysis of attrition rates across the task suggested that participants who were provided a choice were significantly more likely (88.71 percent) to complete the task than those who were not provided a choice (45.16 percent), χ^2 (26.56, n = 124) < 0.001 (see Appendix 18: Study Set 2, Experiments 1-2, Task Engagement).

Digging Deeper: Possible Drivers Behind the Studies' Findings

Drivers of Human-AI Reactance

In the first set of experiments, which address RQ1, participants were asked to make production decisions for several artificial film posters with the goal of making decisions consistent with the majority of other respondents. The task was designed as a proxy for IC decision settings in which end users of AI agents must decide whether to accept an analytic product or recommendation regarding a product. In a decisionmaking setting, participants were provided an AI agent, *Artemis*, that offered them a decision recommendation in either verbal or numeric form. Participants were asked to make these decisions in a series of "practice" and "actual" tasks, and the format of the decision recommendation (i.e., verbal or numeric) was varied randomly across the different sets of tasks depending on participant condition assignment. Thus, format served as a cognate for *User Interface Settings.* The participants determined the extent to which they relied on the AI agent recommendations, and the dependent variable measured the number of production decisions consistent with the AI agent recommendations.

As noted in the Findings section above, the results in both experiments under Study Set 1 were broadly consistent with H1 (Experts will be less susceptible than Nonexperts to deviations in *User Interface Settings*). The results of Study Set 1's experiments showed that Experts were generally not reactive to the consistency of recommendation presentation across the practice and actual tasks. On the other hand, Nonexperts who received incongruent recommendations across the practice and actual tasks (a proxy variable for an environmental variable, *User Interface Settings*) were significantly more likely to concur with the AI agent's recommendations. Furthermore, the pattern of decisionmaking results was explained by a mediating variable, self-assessed Reliance on the AI agent, suggesting the decision to rely on the AI agent was an active one.

Notably, while AI agent recommendations in Experiment 1 were only positively valenced (e.g., framed in terms of a recommendation to produce the film such as "there is a 75-percent probability/it is likely this film would do well"), recommendations in Experiment 2 were more realistic and, therefore, were both positively and negatively valenced. Although the inclusion of both positively and negatively valenced recommendations somewhat attenuated the previously observed differences between observed means in Experiment 2, this result was not unexpected. Increasing the realism of the task necessarily increased the likelihood that respondents might make different decisions based on random,

unobservable factors. Importantly, Nonexperts who received AI recommendations in an Incongruent format remained significantly more likely to accept AI agent recommendations. Furthermore, the observed pattern's dependence on participants' self-assessed Reliance on the AI agent revealed that the participants were consciously aware of their active reliance on the AI agent. Note that the measures used to capture participant reliance on the AI agent were post-task questions (see Appendix 11: Study Set 1, Experiment 2, Reliance Measures and Mediation Results), which suggests—significantly—that participant reliance on the AI recommendations was above the perceptual threshold and, therefore, a function of active participant decisionmaking.

Interestingly, although the pattern of results for Experts generally aligned with preexperiment expectations, the pattern for Nonexperts ran contrary to expectation. As discussed previously, Experts tend to recognize their way through a problem, whereas Nonexperts tend to reason their way through one. Experts have been shown to be better at recalling key information and perceiving subtle differences important to a task. They tend to employ faster and more uniform decisionmaking, which is consistent with employing crystallized knowledge inherent in domain expertise.[82, 83, 84]

On the other hand, the Nonexperts in Study Set 1 presumably reasoned their way through the decision tasks and, when they were exposed to the Incongruent *User Interface Settings*, adjusted their decisionmaking in a pattern generally more consistent with the AI agent's recommendation. Intuitively, one would expect perceived inconsistencies (i.e., incongruencies in recommendation presentation) between practice and actual tasks to be interpreted as an error, and previous research has shown that, when people observe an AI agent make a mistake, they are less likely to accept its recommendations.[85] The opposite pattern was observed.

A possible explanation for the Nonexperts' readiness to accept the AI agent's recommendations, even when perceiving inconsistencies, is the inclination toward uncertainty absorption among Nonexperts. Under uncertainty absorption, the Nonexperts may have felt they did not possess a sufficient understanding of the intricacies of the task, and this might lead to suboptimal performance. Furthermore, the incongruent presentation of AI recommendations in the Nonexpert-Incongruent condition may have served to simultaneously increase the salience of the AI recommendation. Participants may then have shown increased propensity to rely on the AI agent's recommendations because they felt they could not do well on the task with their own level of understanding.[86]

If the observed pattern of results was a function of uncertainty absorption among Nonexperts and relatively crystallized knowledge related to information processing among Experts, one would expect to observe this in their self-assessments of the degree to which they relied on AI agent recommendations—Self-Assessed AI Agent Reliance. Consistent with this possible explanation, the results pattern for both Experiments 1 and 2 in Study Set 1 generally conformed to the results pattern expected under uncertainty absorption: Nonexperts indicated that they relied on the AI agent recommendations significantly more when receiving recommendations with *User Interface Settings* that were incongruent across the practice and actual tasks relative to those assigned to the Expert condition. Additional research is necessary to isolate whether the observed mediation pattern, in which Nonexperts (but not Experts) indicated greater reliance on the AI

agent, is actually because of uncertainty absorption. Thus, the following propositions, consistent with the results pattern, are offered:

P1: **Demonstrated**: The pattern in which Nonexperts (Experts) made decisions consistent (inconsistent) with the AI agent's recvommendations when *User Interface Settings* were Incongruous is mediated by an active reliance on the AI agent's recommendations.

And:

P2: **Proposed**: The pattern in which Nonexperts (Experts) made decisions consistent (inconsistent) with the AI agent's recommendations when *User Interface Settings* were Incongruous is explained by both uncertainty absorption and active reliance on the AI agent's recommendations, in that order.

Improving Joint Human-AI Decisionmaking

The second set of experiments addressed the need to develop additional tools to improve joint human-AI decisionmaking (RQ2). Specifically, Study Set 2 investigated the potential benefit of providing participants a *Choice* in *Explainability Level* of an AI agent's recommendation output, with the reasoning that allowing participants a *Choice* in *Explainability Level* might improve participant task engagement and thus improve overall performance. The first experiment was GEOINT-focused and relied on a sample of IC GEOINT analysts who were presented with a series of images in which they were asked to count the number of four-door sedans. The analysts were assisted by an AI agent that provided recommendations output in either Low or High *Explainability Levels*. The second experiment was similar in overall design although it was an editing-focused task relying on a sample of IC business analysts and editors. The participants were asked to count the number of spelling and grammatical errors in a series of text documents, and the front end of the experiment was altered to provide some participants the opportunity to choose which *Explainability Level* they preferred to receive in the task.

Explainability Level (Study Set 2, Experiment 1)

Consistent with H2 (A High (Low) Explainability recommendation from an AI agent will result in increased (decreased) accuracy), results from Experiment 1 generally showed participants performed more accurately when receiving recommendations in the High *Explainability Level* format. Interestingly, there was no difference in participant engagement in the task between those assigned to the High and Low *Explainability Level* conditions (H4). And although participants assigned to the Low *Explainability Level* condition perceived they relied on the AI agent significantly more (H3a), in fact an analysis of *Actual AI Reliance* revealed no significant differences in *Actual AI Reliance* between the High and Low *Explainability Level* conditions (H3b). Taken together, the improvement in performance accuracy along with the lack of difference in task engagement suggests that provision of output in a High *Explainability Level* format may result in higher participant cognitive performance than when output is provided in a Low Explainability format. However, the observed lack of difference in engagement was not fully consistent with XAI objectives.

Consistent with H5 (A High (Low) Explainability recommendation from an AI agent will result in less (more) dislike toward the AI agent), higher Explainability resulted in significantly decreased dislike toward the AI agent, and this in turn resulted in significantly higher self-perceived (although not actual) reliance on the AI agent.

Note that this result contrasts with the participants' self-perception of reliance cause and effect: although participants perceived they relied more on the AI agent in the High *Explainability Level* condition, they also perceived that they did so with increased dislike toward the AI agent. Thus, while overall performance was better for the High *Explainability Level* condition, there was a strong emotional preference for the Low *Explainability Level* condition.

Previous research has shown that, in the long run, people tend to do better at tasks they enjoy and avoid interactions they code as negatively valenced.[87] Furthermore, they tend also to have a strong preference for interactions with less complexity. [88, 89, 90] People tend to prefer simpler causal explanations and simpler versions of concepts, for example, and although consumers are generally willing to pay more for products with more features, there is a point at which more is actually less. [91, 92, 93] A challenge with increasing *Explainability Level* is that it is frequently (although not always) associated with presenting additional information that may be perceived as more complex. The participants' exhibited dislike for the High *Explainability Level* algorithm is generally consistent with aversive behavior toward more complex interactions and with coding the experience with a negative valence.

A further consideration is that mere explainability may not be enough to improve end users' willingness to use the most sophisticated algorithms. More sophisticated AI algorithms have been characterized as explanatory "black boxes" because even their designers may not fully understand the reasoning behind their recommendations. For example, neural network algorithms model the underlying data by varying relationships between nodes of different weights spread across a variable number of layers. Although it is certainly possible to characterize the arrangement, weights, and relationships between the nodes, it is not clear from a human perspective—even for the algorithm's developers—what any particular arrangement, set of relationships, or vector of nodal weights might mean. Thus, the overall algorithmic procedure is clear, but the way in which the algorithm actually models the underlying data—the reasoning behind the ultimate recommendation—is not. Therefore, consistent with the results for H5, providing such explanations may further increase the complexity of High Explainability results, subsequently increasing dislike toward the AI agent and decreasing self-perceived reliance on it.

Thus, end users may have cause to dislike AI agents providing High Explainability output because of the often necessarily increased complexity associated with the explanations. While previous research has shown that people may be generally more receptive toward AI output that is more explainable, most studies that support these results only assess increased explainability using relatively simple algorithms rather than those consistent with being a "black box."

And although people may adapt to the increased complexity usually associated with the High *Explainability Level* results of more complex algorithms, existing research shows that—for at least some populations—aversion

to computationally intensive activities, such as those involving algorithms, may increase rather than decrease over time.[94] More research in this area is warranted. Thus, even while real benefits are observed that are consistent with the XAI thesis, these gains may be short-lived if people increasingly avoid AI agents as a function of increased dislike because of relatively increased complexity.

Choice of Explainability Level (Study Set 2, Experiment 2)

In Experiment 1 the results provided qualified support for the XAI thesis. Although performance accuracy improved after receiving AI agent recommendations in a High Explainability format and although participants *perceived* they relied on the AI agent more (albeit in the Low Explainability condition), they did not *actually* rely on the AI agent more. Moreover, task engagement did not improve. In contrast to this, in Experiment 2 participants demonstrated significantly greater task engagement when they were offered a *Choice* in the AI agent's *Explainability Level*. Furthermore, participants who were offered a *Choice* in the AI agent's recommendation *Explainability Level* performed as well in both the Low and High Explainability conditions as in the No Choice/High *Explainability Level* condition. In other words, the observed performance accuracy advantage gap between High and Low *Explainability Levels* fully attenuated in favor of increased accuracy when participants were offered a Choice.

Table 4: Summary of Findings by Hypothesis

H1	Experts will be less susceptible than Nonexperts to deviations in *User Interface Settings*.	**Supported**
H2	A High (Low) Explainability recommendation from an AI agent will result in increased (decreased) accuracy.	**Supported**
H3a	Participants will perceive they rely on the AI agent more (less) in the High (Low) Explainability condition.	**Not supported: opposite pattern**
H3b	Participants will actively rely on the AI agent more (less) in the High (Low) Explainability condition.	**Not supported**
H4	A High (Low) Explainability recommendation from an AI agent will result in more (less) task engagement.	**Not supported**
H5	A High (Low) Explainability recommendation from an AI agent will result in less (more) dislike toward the AI agent.	**Supported**
H6	Allowing participants a choice (no choice) in Explainability Level format will result in increased (decreased) task engagement.	**Supported**
H7	Allowing participants a choice (no choice) in Explainability Level format will result in increased (decreased) accuracy.	**Supported**
H8	Allowing participants a choice (no choice) in Explainability Level format will result in decreased (increased) disliking irrespective of Explainability level.	**Partially supported**

Interestingly, despite the apparent improvement in performance accuracy, participants still did not *actually* rely on the AI recommendation more in either the *Choice* or *Explainability Level* conditions. However,

perceived reliance on the Low *Explainability Level* AI agent recommendations persisted. Since participants' improvement in accuracy cannot be attributed to actual increased reliance on the AI recommendations, this suggests the observed performance was a result of participants working harder on the task. Interestingly, the previously observed dislike toward the AI agent in the High Recommendation formats also attenuated, which is encouraging from a long-term implementation perspective. As previously discussed, there is reason to be concerned that interactions with the AI agent that are coded as negatively valenced may result in decreased usage over the long run. However, it is unclear from this experimental setup whether the attenuation of dislike was a function of different task domains (i.e., GEOINT vs. editing) or the knowledge that there were different *Explainability Levels* available.

Limitations and Future Research

There are several shortcomings to this research. First, Study Set 1 employed a highly contrived task with which participants were largely unfamiliar (film production decisions). Although the participants were led to believe they had relatively more or less expertise in the task domain, the fact remains that knowledge in this area was artificial: the proposed films were not real, the films were never going to be produced, participant knowledge was based on contrived (and manipulated) information sets, and participants did not exercise any real authority over whether the films would be produced. Thus, despite the fact that participant decisions were consequential from a task incentive perspective, one could reasonably argue that participants might have been making choices consistent with what the researcher intended. If so, the observed pattern of results might reflect what participants believed the researcher wanted to see rather than their preferred decisions.

Second, recall that Self-Assessed Expertise was a manipulated variable based on the participants randomly receiving one of two different information sets and their subsequent performance on a test. In real life, expertise takes years of focused, disciplined study. Although an oversimplification, research has shown that people need more than 10,000 hours of experience in a given domain before they are "experts" and exhibit qualities consistent with domain expertise.[95] While the psychological manipulations of expertise in these studies were successful—in that people *believed* they were experts—it is possible that the artificiality of the experiment failed to activate a mindset utilizing cognitive processes consistent with actual expertise. If so, the observed pattern of results may have resulted from an unobserved, yet-undetermined variable. Additional exploration of these results in a field setting involving actual experts is warranted as an important next step.

Finally, there are practical concerns. While this research shows that Experts and Nonexperts are differently susceptible to AI agent recommendations when receiving recommendations with *User Interface Settings* that are Incongruent between Practice and Actual Tasks, follow-on options for managers and developers are somewhat limited. For example, with respect to differences in expertise levels, managers may attend more carefully to shift assignments and how analysts with relatively greater domain expertise are integrated into their workforce. And developers and managers alike may develop and implement policies to test and

standardize *User Interface Settings* across systems to mitigate the potential for the effects observed in Study Set 1's experiments. However, in both cases, managers and developers are simply reducing the differential influence of the AI agent on known population segments with now-identified environmental characteristics (*User Interface Settings*). The research does little to assess and recommend improvements for the overall performance of the human-machine team.

Study Set 2 addressed the above concerns and presented results that focused on improving joint human-AI decisionmaking. Specifically, participants were members of the IC and performed tasks relevant to their stated area of domain expertise. Be that as it may, there are several limitations to this research that bear further investigation.

First, it is interesting that provision of mere *Choice* improved task engagement but *Explainability Level* did not. Further exploration of the role of *Choice*, potentially at other touchpoints in the human-AI interaction lifecycle, is warranted. For example, one might explore allowing customization of specific aspects of the algorithm, such as distance measurement, or of elements of the algorithmic model, or one might even allow selection of different types of algorithms. Understanding how the benefits of choice generalize across the human-AI interaction lifecycle will allow mission owners to better customize when and where to build in choices for end users.

Second, while *Choice* appears to have a significant main effect on task engagement and a significant joint interactive effect with *Explainability Level* on accuracy, note that a significant body of research indicates that provision of too much choice can impair decisionmaking.[96] As mentioned above, this research only considered two levels of *Choice* with respect to one human-AI touchpoint (*Explainability Level*). The number of choice options can quickly expand depending on the number of human-AI touchpoints allowing choice, as well as the number of choices offered. Understanding the optimal number of touchpoints to allow *Choice* and how many choices to allow at each touchpoint is critical.

Third, future research should focus on better understanding the discontinuity between actual and perceived reliance on the AI agent recommendations. One possible explanation for this discontinuity might be the timing of when the measures for actual and perceived reliance were collected. Recall that actual reliance was captured during the task, whereas perceived reliance was captured post-task with a set of behavioral survey items. Previous research has shown that momentary assessments, such as impressions, during a task may differ significantly from global or retrospective assessments post-task.[97, 98] Momentary assessments have been shown to better predict impulsive behaviors, whereas retrospective assessments tend to better predict behaviors in which cognition (and by necessity memory) plays a role.[99, 100] Study Set 2's experiments were both cognitively focused (i.e., involving image recognition, problem solving, and counting). Additional research should explore whether the observed results pattern differs for tasks involving impulsive or "spur of the moment" decisionmaking.

Finally, it is unclear why mere awareness of *Explainability Level* options in Experiment 2 resulted in attenuation of the previously observed mediation pattern for dislike of the AI agent. If a High *Explainability Level* was a driver for dislike, then awareness of other *Explainability Level* options should not have attenuated the

mediation results. Furthermore, presenting additional *Explainability Level* options necessarily *increased* the complexity of the overall interaction, and one would intuitively expect that increased complexity would invite increased dislike. This pattern was also not observed. Further research investigating potential interactive influences of option awareness on *Choice*, *Explainability Level*, and other potential human-AI touchpoints is warranted.

Conclusion: Implications and Recommendations

Human-AI Relationship Research and Results

IC and DoD investments in AI systems in analytic settings have produced mixed results. Although there have been some successes with limited human-AI integration, more complex applications have proven elusive. In particular, some of the challenges discussed in this work include: end users who are unwilling to accept AI agent recommendations; the potential introduction, perpetuation, and accentuation of human bias in human-in-the-loop human-AI systems; and in the research and management literature, an overweighting of AI agent recommendation performance (i.e., number of false positives and negatives) and an underweighting of joint human-AI outcomes.

The goal for this research has been to draw attention to the importance of determining and addressing specific needs of the human element in the human-AI relationship when developing complex algorithms for integration in IC missions. The author has demonstrated this through the investigation of two research objectives: identification of an IC-relevant driver of human-AI reactance—broadly defined as an irrational avoidance or attraction to AI agents—that is managerially relevant (RQ1), and identification of additional approaches to improving joint human-AI decisionmaking (RQ2).

With the goal of addressing RQ1, Study Set 1 investigated the interactive role between an environmental factor (*User Interface Settings*) and a psychological factor (*Expertise Level*)—both of which are common to IC settings and have not been fully accounted for in the extant literature. Results showed that while Experts were relatively insulated from the influence of different *User Interface Settings* on their decisionmaking, Nonexperts were not, and when Nonexperts encountered a situation with which they were unfamiliar (Incongruent *User Interface Settings*) they were significantly more likely to rely on the AI agent's recommendation. While this may be a positive outcome depending on the relative accuracy of the AI agent and the human analyst, the extent to which this happens differently between Experts and Nonexperts may be problematic for mission owners, who do not wish to see the consistency of decisionmaking depend on random shift assignments and subsequent differences between shift compositions of Experts and Nonexperts. One area that was not addressed in Study Set 1 was quality of decisionmaking, or the extent to which joint human-AI reasoning produced improved outcomes.

Study Set 2, which was designed to address RQ2, further investigated the influence of *Explainability Level* and *Choice* in improving joint human-AI decisionmaking. Results pertaining to *Explainability*

Level suggested that, although increasing *Explainability Level* resulted in improved performance accuracy, this also came at the expense of increased participant dislike for the AI agent. Furthermore, High and Low *Explainability Level* did not significantly predict task engagement which is a key objective of the XAI thesis. A further investigation into allowing participants *Choice* in AI agent *Explainability Level* revealed that participants who were allowed *Choice* performed as accurately as participants in the previously observed No Choice/High Explainability condition. Furthermore, allowance of *Choice* simultaneously improved overall task engagement and decreased participant dislike toward the AI agent. Taken together, these results suggest previously observed benefits to increased *Explainability Level* may at least partially be an artifact of tightly controlled experimental designs, and subsequently may not generalize well to settings in which participants realize they have a choice in recommendation output—a situation that has become increasingly common given the recent trend toward integration of analysts with data scientists and developers.

Implications and Recommendations

These findings suggest five recommendations to improve integration of AI agents with human analytic efforts:

1. IC managers should design analytic team assignments to ensure optimal human-AI interoperability.

When designing and integrating AI agents that provide recommendations to analysts, managers should attend to the distribution of domain experts across workgroup assignments. Findings from Study Set 1 showed that differences in domain expertise resulted in significant differences in the degree to which AI recommendations were accepted. These findings further depended on *User Interface Settings* applied to Non-experts but not Experts. These results suggest that work groups with differences in the depth of expertise in their bench strength may arrive at different conclusions.

Currently, analytic teams are staffed based on need and according to a billet structure. Such workgroups tend to be pyramid-shaped, with few experts providing analytic and managerial guidance to a relatively greater number of less-experienced analysts. Given the relatively small size of most offices, this assignment pattern may lead to the erroneous belief that expertise levels are randomly distributed. However, in practice, the distribution of expertise may be unbalanced.[101] This pyramid-shaped structure can result, therefore, in significant differences in the number of experts assigned across workgroups and shifts, leading to differences in the extent to which AI agent recommendations will be accepted.

2. IC managers and AI system developers should routinely monitor how the system's user interface designs contribute to different decisionmaking outcomes.

Where it is not possible to manage the distribution of domain expertise across analytic teams, managers may wish to identify additional user interface design factors that contribute to differences in recommendation acceptance rates. Further, managers should work with AI system developers to ensure that these factors do not skew decisionmaking outcomes across workgroups with differing levels of expertise.

This research has shown generally little difference in whether participants accepted AI agent recommendations when they were framed in either a numeric or verbal format. However, participants did recognize when recommendation format changed over time (i.e., numeric-verbal or verbal-numeric vs. numeric-numeric or verbal-verbal), which resulted in different AI agent recommendation acceptance rates between domain Experts and Nonexperts, who were significantly more likely than the Experts to rely on the AI agent. Failure to be aware of this or control for it may result in a lack of analytic decisionmaking consistency across workgroups.

3. IC managers should identify and mitigate the effects of additional individual difference factors that may influence decisionmaking outcomes.

This monograph has focused on the interactive role of domain expertise on certain user interface settings such as verbal or numeric presentation of data. However, the scope of the research was necessarily limited by pragmatic factors such as time and cost to implement various experimental designs. Additional individual differences that interact with the user interface settings explored, as well as additional user interface settings, should also be considered. Managers should continually scan for additional individual difference factors that may significantly influence end user decisionmaking.

After identifying such differences, managers would also be wise to develop education and training campaigns to minimize gaps across analytic teams. For example, one workgroup may possess relatively more individuals with greater expertise in Order of Battle assessments, whereas another workgroup possesses relatively more individuals with greater expertise in C4ISR. In such cases, it may be reasonable to provide crosstraining and education opportunities to balance the distribution of expertise.

Last, it may not be possible to control for all these factors. For example, some individual differences that result in disparate AI acceptance rates may fall into protected categories. Education campaigns for both analysts and supervisors can help guard against specific biases linked to these individual differences.

4. Explore and identify the touchpoints between human and AI interactions in which choice should be offered and determine optimal choice structure.

This monograph has shown that providing end users a choice regarding *Explainability Level* led to increased engagement and subsequently more accurate performance—despite no difference in the amount of time spent on the task. However, there are multiple touchpoints between humans and AI agents that should be further explored. For example, offering humans the option to help select model parameters could yield generally improved insight into the "black box" and subsequently increase overall performance. Alternately, offering humans the option to select different forms of avatars that present information to the end user, while seemingly innocuous, may result in significant differences in engagement and overall performance.

Furthermore, research into choice has shown that more choice is not always better. Customers can easily suffer from "choice overload" and this can lead to suboptimal outcomes as a function of constructed preferences.[102] Thus, it is critical that managers not only systematically catalog various touchpoint opportunities

between human and AI agents but also optimize them for the ideal set of choice options. Furthermore, managers must test these touchpoints with the goal of mitigating negative influences on decisionmaking while also capitalizing on the positive influences.

5. **Finally, IC managers must focus on joint human-AI outcomes in their implementation of AI among the analytic community.**

Too often, assessments of algorithm performance have been framed in terms of model fit criteria or outcome performance based on the number of false negatives and positives. For example, a supervised machine learning model might be trained on a training data set comprising 80 percent of recorded data and human decisions, and then tested against a test data set comprising 20 percent of recorded data and human decisions to assess model performance relative to the human decisions. Models that perform as well in the 20-percent test set as in the 80-percent training set are determined to be "good." Alternately, algorithmic decision accuracy may be determined based on known outcomes to develop a confusion matrix providing the ratios of false positives, false negatives, true positives, and true negatives. Relatively greater ratios of true positives and negatives to false positives and negatives result in an assessment that the model is performing as expected.

Taken together, however, the findings from this research suggest that, in joint AI-human decisionmaking systems (such as those with a human-in-the-loop), the mere introduction of algorithmic decisionmaking aids or recommendations can influence not only algorithm aversion or appreciation—as recently suggested by research from Berkeley Dietvorst[103] and Jennifer Logg[104]—but also the actual quality of the joint decision in a manner inconsistent with the original intent. Thus, developers, managers, and researchers in this area should take care to ensure that joint decisionmaking does not negatively influence overall performance or decisionmaking speed. Furthermore, while this research considers the impact of joint decisionmaking in the short term, managers and future researchers should also consider the long-term effects on humans of ceding critical thinking to AI-based agents.

Appendix 1: Study Set 1, Overview of Task, Instructions, and Incentive Structure

Task Background

We are testing a new artificial intelligence (AI) agent, Artemis, that can assist people with making decisions based on rules or "tips" we have provided it. We want to see how well people work with Artemis.

Task Overview

We have designed a task in which we ask you to play the role of film producer. We will show you an image of a (fake) movie poster and are interested in whether you would either BEGIN PRODUCTION or CANCEL PRODUCTION based on the visual appeal of the movie poster. Artemis may also provide you a recommendation and, if you receive this, you may use this as you see fit.

BONUS OPPORTUNITY

If you select the choice (BEGIN PRODUCTION or CANCEL PRODUCTION) MOST frequently selected by previous participants, you will GAIN $0.05 for each correct choice.

However, if you select the choice (BEGIN PRODUCTION or CANCEL PRODUCTION) LESS frequently selected by previous participants, you will LOSE $0.05 for each incorrect choice.

Please ensure you understand the above instructions and proceed to the next screen to give your informed consent.

Appendix 2: Study Set 1, Information Sets/Tips

Before you begin, we need to assess your current level of expertise related to the task (film production decisions). Our previous work has shown that expertise level can influence your answers, and we need to control for this in our analysis.

You may not think you're an expert in film production decisions, but in fact research has shown that everyone varies in their understanding of different topics, and some people are more or less expert than others. Research has shown that people with more expertise in an area are better able to process and use information related to that topic. Next, we will give you some tips that will help you make your decisions in the actual task, and then we will give you a comprehension test of the tips that we will give you. Your score will determine your relative level of expertise.

On the next screen you will see the tips related to film production and the movie posters we are about to show you. Please read these carefully and respond to the five comprehension questions as best you can. We will score your answers and provide you with feedback on both your answers (if incorrect) as well as your level of expertise relative to the average.

Bonus Opportunity: Because our analysis depends on successfully controlling for your prior expertise in this task, we are incentivizing this section. For every answer you score correctly you will earn an additional $0.05. Incorrect answers will receive no additional incentive.

Click the --> button to begin the assessment.

Tips on Selecting Films for Production Based on the Poster:

1. Most people have heard of how the "rule of thirds" applies to photography. The rule refers to the phenomenon that people find photographs more visually appealing when the photographed area is divided (by the subject, text, etc.) into clearly visible thirds. This also applies to film posters. However, because film is narrative and dynamic whereas photographs are static, people expect to see the thirds arranged to be consistent with how stories are written: that is, top-to-bottom when vertical, and left-to-right when horizontal.
2. When advertising for an upcoming film, artists are also careful to take into account both where and how main characters are placed in the image. Because people expect stories to naturally run

from left-to-right or top-to-bottom, they also expect the protagonists or "good guys" to be on the left side of the poster facing right (congruent with the story's direction), whereas they expect the antagonists or "bad guys" to be on the right side of the poster facing left (incongruent with the story's direction).

3. Sometimes, when movie producers wish to subconsciously signal a significant plot reversal (e.g., the protagonist suddenly turns out to be the antagonist, and the antagonist suddenly turns out to be the protagonist) it helps to flip the expected subject's placement and orientation in the image. Artists may also break the "rule of thirds" to highlight this. This can be especially effective when employed correctly (i.e., in a film in which the plot reverses) but can backfire when employed incorrectly (i.e., when the film's plot does not reverse).

Appendix 3: Study Set 1, Attention Check

1. If a film has two main characters (a boy and a girl), what would be the optimal placement for the characters in the film poster to generate the most interest:

    - Girl on the left facing right; Boy on the right facing left
    - Girl on the top-left facing downward to the right; Boy on the Top-right facing downward to the left
    - Boy on the right facing left; Girl on the left facing right
    - Boy on the top-left facing downward to the right; Girl on the top-right facing downward to the left
    - Boy and girl side-by-side facing the "camera" or observer directly

2. If a film has two main characters (a man and a woman), what would be the optimal placement for the characters in the film poster to generate the most interest:

    - Woman on the left facing right; Man on the right facing left
    - Woman on the top-left facing downward to the right; Man on the top-right facing downward to the left
    - Man on the right facing left; Woman on the left facing right
    - Man on the top-left facing downward to the right; Woman on the top-right facing downward to the left
    - Man and Woman side-by-side facing the "camera" or observer directly

3. In a conventional story, which main character type would be best placed centered, looking down in the top half of the film poster?

    - A protagonist man
    - An antagonist woman
    - A protagonist girl
    - An antagonist boy
    - A protagonist child (could be a boy or girl)

4. In a conventional story, which main character type would be best placed centered, facing right in the left third of the film poster?

    - A protagonist animal
    - A protagonist man
    - A protagonist woman
    - A protagonist girl
    - A protagonist boy

5. In a story with a plot reversal, which main character type would be best centered, facing left in the right third of the film poster.

    - An antagonist animal
    - An antagonist man
    - An antagonist boy
    - An antagonist woman
    - An antagonist girl

Appendix 4: Study Set 1, Sample AI Agent Recommendations Following Tutorial

Source: Movie poster concept derived from reddit site /u/Your_Post_As_A_Movie. Images used to create this image from Pexels.

The AI Agent (*Artemis*) recommendations were presented in semantically congruent or incongruent format, depending upon whether an individual's tutorial had used a verbal or numeric format.

If the tutorial practice task was worded in verbal format, as "*Artemis* suggests it is likely this film would do well if you begin production," then:

Congruent format:

Participants Shown: "*Artemis* suggests it is likely this film would do well if you begin production."

OR

Incongruent format:

Participants Shown: "*Artemis* suggests there is a 75-percent probability this film would do well if you begin production."

If, on the other hand, the tutorial practice task was worded in numerical format, as "*Artemis* suggests here is a 75-percent probability this film would do well if you begin production," then:

Congruent format:

Participants Shown: "*Artemis* suggests there is a 75-percent probability this film would do well if you begin production."

OR

Incongruent format:

Participants Shown: "*Artemis* suggests it is likely this film would do well if you begin production."

Appendix 5: Study Set 1, Self-Assessed Expertise

I am an Expert at discerning film posters for production. [Likert Scale, 1-7, Disagree to Agree]

I am better than average at discerning film posters for film production. [Likert Scale, 1-7, Disagree to Agree]

I am good at discerning film posters for film production. [Likert Scale, 1-7, Disagree to Agree]

Variable	Description	Experiment 1	Experiment 2
Self-Assessed Expertise	Expert	5.14 (0.19)	5.38 (0.15)
	Nonexpert	1.60 (0.18)	1.97 (0.16)

Appendix 6: Study Set 1, Demographics

		Distribution of Survey Responses		
Variable	**Description**	**Diagnostic Test**	**Experiment 1**	**Experiment 2**
Gender	Male	44.14%	47.57%	41.40%
	Female	53.15%	51.46%	57.96%
	Other	2.70%	0.97%	0.64%
Age	Age group of participants:			
	18-25	18.92%	9.71%	13.38%
	26-35	41.44%	30.10%	39.49%
	36-45	20.72%	29.13%	26.11%
	46-55	7.21%	22.33%	13.38%
	46-65	9.01%	3.88%	5.10%
	66 and above	2.70%	4.85%	2.55%

Appendix 7: Study Set 2, Experiment 1, Study Instructions

Introduction

Thank you for agreeing to participate in this survey. Pretesting has shown that the total time for this survey ranges between 10 and 20 minutes depending on how long you choose to spend on each of the tasks.

Please complete this survey using a laptop, desktop, or tablet computer with a reasonably large screen. It is nearly impossible to complete using a mobile/smartphone.

The instruction pages require you to spend a minimum of 20 seconds before you can proceed to the next page. Once the 20 seconds have elapsed an arrow --> will appear such as the one you see in the lower-right corner of the page.

It is extremely important that you take this survey without assistance from others. We want your individual responses. Also, it is also important that you not share your experience or answers with others. Sharing will undermine the validity of the study. Please acknowledge you will answer the survey without information or assistance from others.

Task Background

Counting (e.g., people, cars, etc.) is a tedious but necessary task many analysts perform at some point in their careers. We have developed an artificial intelligence (AI) agent to assist with this. The idea is that, if an AI agent can count entities for you, this frees you to perform more interesting tasks. The agent we designed employs the latest algorithms, and if widely available would be a significant advancement over existing tools. Specifically, in this task we are testing how the AI agent performs when counting a type of vehicle (four-door sedans), although it could be used to count anything once properly calibrated.

Task Overview

We will show you four static images, each with several cars either parked or in motion. We want you to count the number of four-door sedans in the image. Feel free to use scratch paper if you need it to help keep track. Once you complete your count you will be asked to provide your assessed number of four-door sedans on the following page.

During the task our AI agent will also assess the number of four-door sedans using its algorithm. The algorithm itself is actually comprised of six different models, and relies on training data regarding vehicles as well as various contextual clues. The AI agent does not know the actual number of four-door sedans, and is working with the same information that you have. After the AI agent has finished calculating, it will provide its recommendation, and you may use this information in your own assessment if you wish. We will ask you questions at the end of the survey that capture whether you relied on the AI agent or not. Again, it is your choice whether you rely on the AI agent's assessment or not.

You will not be paid for this task. However, at the conclusion of our research we will make available a copy of any resulting publication. Note that all your responses will be anonymous from our perspective.

Target Examples

On the next page we will show you a picture of what you will be asked to count.

Below is an example of a four-door sedan. We are asking you to count how many of these you identify in each image.

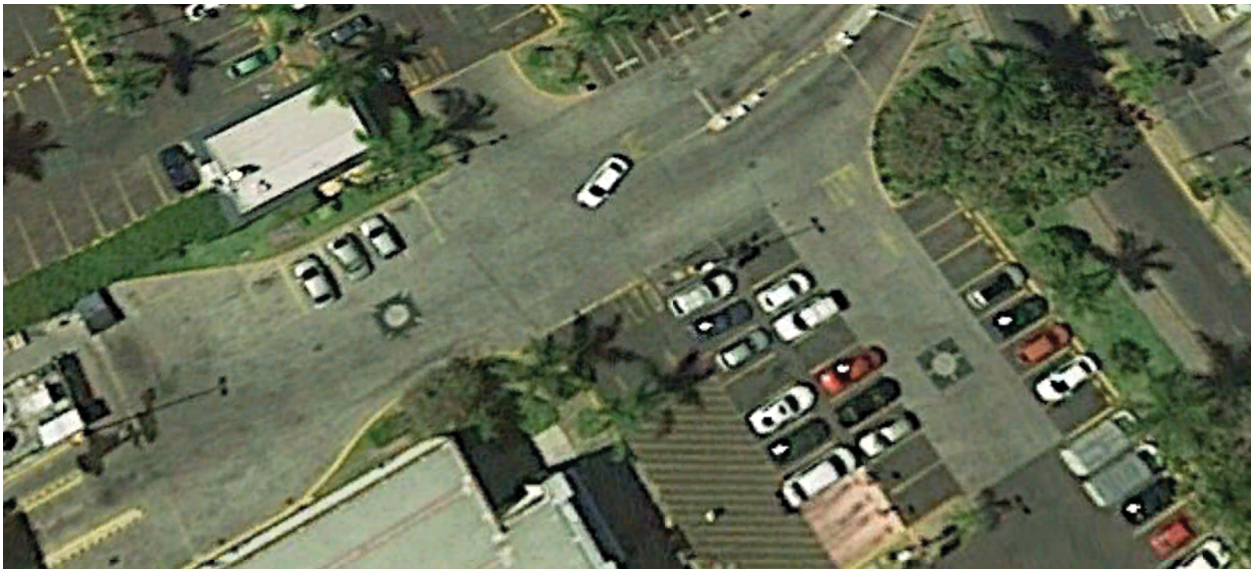
Four-door sedan:

However, note there are also two-door sedans (shown below), hatchbacks of two- (not shown) and four-door variety (shown below), as well as trucks (not shown), vans (not shown), and other vehicle types (not shown). **Be mindful to count only vehicles you think are four-door sedans.** Take a moment now to familiarize yourself with the similarities and differences.

Two-door sedan:

Four-door hatchback:

Below, is an example of the type of image you will see. Remember, in this study we wish to know how many four-door sedans you can find.



Source: "Parking lot of Restaurante Mar y Tierra Veleiros, Jalisco, Mexico," Google Maps, accessed on March 14, 2021, https://www.google.com/maps/@20.709417,-103.41092,45m/data=!3m1!1e3.

At the same time that you view the image, the AI will also view it and assess the number of four-door sedans using its algorithm. As stated before, the AI agent's algorithm considers a training data set as well as various image-specific contextual factors such as size, shape, and shadow, and calculates six models each with its own recommended number. The AI agent selects the model with the best "fit" to the data, and provides you its own assessment (count of the number of four-door sedans). It is up to you how much you use the AI agent's assessment. We will ask you questions at the end of the survey to capture how much you chose to rely on it.

Acknowledge you understand the instructions and click --> to advance to the next page and see this in action.

Appendix 8: Study Set 2, Self-Assessed AI Reliance

$\alpha = 0.84$

Rely A: I relied on the AI Agent's recommendations. [Likert Scale, 1-7, Disagree to Agree]

Rely B: It is a good idea to use the information provided by the AI agent in your own assessment. [Likert Scale, 1-7, Disagree to Agree]

Rely C: Someone would do well to rely on the AI agent's recommendations. [Likert Scale, 1-7, Disagree to Agree]

Rely D: The AI agent was a reliable source of information. [Likert Scale, 1-7, Disagree to Agree]

Variable	Description	Experiment 1	Experiment 2
Self-Assessed AI Reliance	Low Complexity	3.72 (0.14)	3.90 (0.13)
	High Complexity	3.35 (0.13)	3.30 (0.13)

Appendix 9: Study Set 2, Participant Demographics and Self-Assessed Expertise

Variable	Description	Experiment 1	Experiment 2
Sample Size	Number of Study Participants	75	124
Gender	Male	63.83%	50.00%
	Female	36.17%	50.00%
	Other	0.00%	0.00%
Age	Average Participant Age	37.77 (10.61)	40.06 (12.18)
Experience	Experience in Assessed Domain (Years)	8.16 (7.28)	14.70 (12.11)
Self-assessed Expertise	Likert Scale, 1-7, Inexperienced to Experienced	5.36 (1.79)	5.00 (2.04)

Appendix 10: Study Set 1, Experiment 1, ANOVA Results (Agreements)

Omnibus Test for Production Decision Predicted by UI Settings

Effect (Type III)	Num DF	Den DF	F Value	Pr > F
Congruency Condition	1	515	6.41	0.012
Expertise Level	1	515	4.72	0.038
Expertise*Congruency	1	515	4.01	0.068

Table of Means

Expertise	UI Settings	Estimate	SE	DF	t Value	Pr > |t|
Congruent	Expert	3.722	0.244	515	15.28	<.001
Congruent	Nonexpert	3.787	0.232	515	16.32	<.001
Incongruent	Expert	3.902	0.261	515	14.94	<.001
Incongruent	Nonexpert	4.905	0.287	515	17.10	<.001

Table of Comparison of Means

Expertise	UI Settings	Expertise	UI Settings	Estimate	SE	DF	t Value	Pr > |t|
Congruent	Expert	Congruent	Nonexpert	-0.065	0.337	515	-0.19	0.848
Congruent	Expert	Incongruent	Expert	-0.180	0.357	515	-0.50	0.615
Congruent	Expert	Incongruent	Nonexpert	-1.183	0.377	515	-3.14	0.002
Congruent	Nonexpert	Incongruent	Expert	-0.115	0.349	515	-0.33	0.743
Congruent	Nonexpert	Incongruent	Nonexpert	-1.118	0.369	515	-3.03	0.003
Incongruent	Expert	Incongruent	Nonexpert	-1.003	0.388	515	-2.59	0.010

Appendix 11: Study Set 1, Experiment 2, Reliance Measures and Mediation Results

Behavioral Questions Regarding Reliance About the Task:

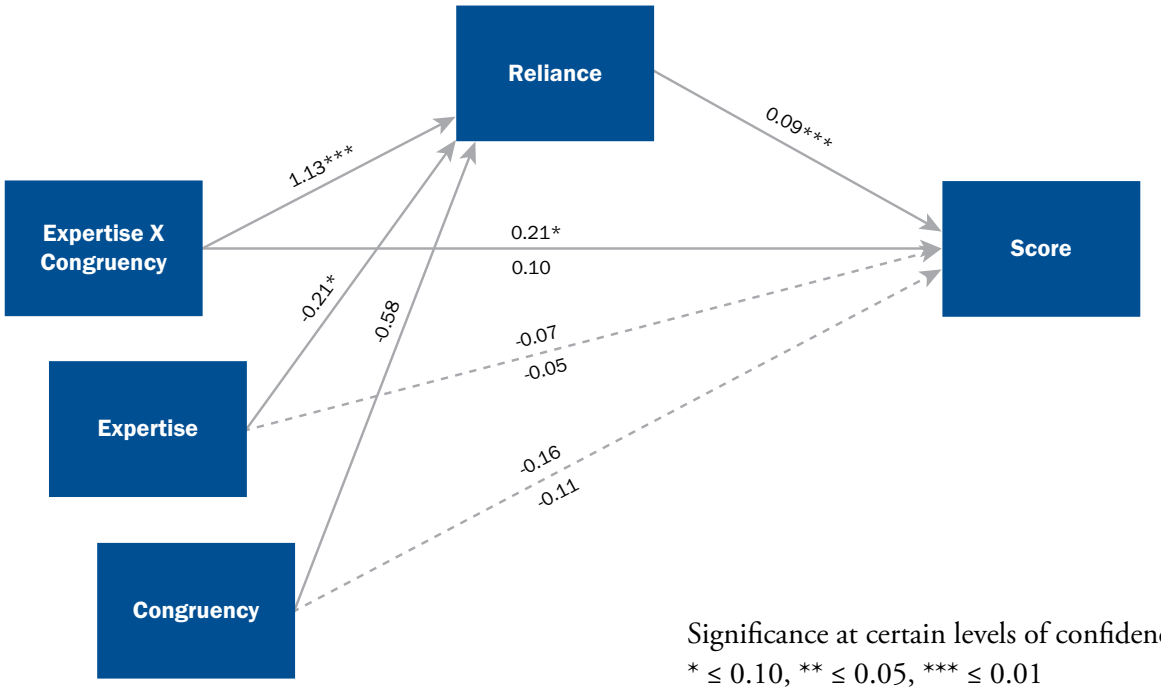
Reliance, α = 0.84

Reliance A: The provided recommendation informed my choice. [Likert Scale, 1-7, Disagree to Agree]

Reliance B: I accepted the provided recommendation. [Likert Scale, 1-7, Disagree to Agree]

Reliance C: The provided recommendation gave me an idea whether to produce the movie. [Likert Scale, 1-7, Disagree to Agree]

Mediation Analysis:



Significance at certain levels of confidence:
* ≤ 0.10, ** ≤ 0.05, *** ≤ 0.01

Appendix 12: Study Set 1, Experiment 2, Self-Confidence Measures

Behavioral Questions Regarding Self-Confidence About the Task:

Confidence, $\alpha = 0.89$

I was good at this task. [Likert Scale, 1-7, Disagree to Agree]

My choices were well-informed. [Likert Scale, 1-7, Disagree to Agree]

I knew what I was doing. [Likert Scale, 1-7, Disagree to Agree]

I was confident about the decisions I made. [Likert Scale, 1-7, Disagree to Agree]

I knew which choices to make. [Likert Scale, 1-7, Disagree to Agree]

Appendix 13: Study Set 1, Experiment 2, ANOVA Results (Agreements)

Omnibus Test

Effect (Type III)	Num DF	Den DF	F Value	Pr > F
UI Settings	1	785	1.23	0.267
Expertise Level	1	785	0.38	0.536
Expertise*UI Settings	1	785	4.00	0.046

Table of Means

Expertise	UI Settings	Estimate	SE	DF	t Value	Pr > |t|		
Expert	Congruent	4.4106	0.2322	785	28.19	<.0001	4.4106	0.2322
Expert	Incongruent	4.2126	0.2051	785	29.53	<.0001	4.2126	0.2051
Nonexpert	Congruent	3.8524	0.2110	785	24.62	<.0001	3.8524	0.2110
Nonexpert	Incongruent	4.5242	0.2280	785	29.95	<.0001	4.5242	0.2280

Table of Comparison of Means

Expertise	UI Settings	Expertise	UI Settings	Estimate	SE	DF	t Value	Pr > |t|
Expert	Congruent	Expert	Incongruent	0.0459	0.072	785	0.64	0.522
Expert	Congruent	Nonexpert	Congruent	0.135	0.0760	785	1.78	0.075
Expert	Congruent	Nonexpert	Incongruent	-0.025	0.073	785	-0.35	0.727
Expert	Incongruent	Nonexpert	Congruent	0.089	0.073	785	1.22	0.223
Expert	Incongruent	Nonexpert	Incongruent	-0.071	0.070	785	-1.02	0.309
Nonexpert	Congruent	Nonexpert	Incongruent	-0.161	0.074	785	-2.16	0.031

Appendix 14: Study Set 2, Experiment 1, ANOVA Results (Accuracy)

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Explainability Level	1	45	5.37	0.0251

Condition*Choice Least Squares Means					
Explainability Level	**DF**	**t Value**	**Pr > |t|**	**Mean**	**SE**
Low Explainability	45	16.11	<.0001	6.06	0.6779
High Explainability	45	9.28	<.0001	3.9432	0.5829

Cohen's d = 3.35

$r_{Y\lambda} = 0.86$

Appendix 15: Study Set 2, Experiment 1, Actual AI Reliance Compared to Self-Assessed AI Reliance

Actual Reliance on the AI Recommendation

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Explainability Level	1	45	1.32	0.2574

Condition Least Squares Means					
Complexity Level (0 = Simple, 1 = Complex) RUN	**DF**	**t Value**	**Pr > |t|**	**Mean**	**SE**
Low Explainability	45	17.4	<.0001	6.01	0.6194
High Explainability	45	13.4	<.0001	5.0114	0.603

Self-Assessed Reliance on the AI Recommendation

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Explainability Level	1	45	3.64	0.0627

Condition Least Squares Means					
Explainability Level	**DF**	**t Value**	**Pr > |t|**	**Mean**	**SE**
Low Explainability	45	35.21	<.0001	3.72	0.1388
High Explainability	45	30.41	<.0001	3.3523	0.1333

Cohen's d = 2.74

$r_{Y\lambda} = 0.81$

Appendix 16: Study Set 2, Experiment 2, ANOVA Results (Accuracy)

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Explainability Level	1	68	7.45	0.0081
Choice	1	68	2.16	0.1459
Explainability Level*Choice	1	68	9.62	0.0028

Condition*Choice Least Squares Means						
Explainability Level	**Choice**	**DF**	**t Value**	**Pr > |t|**	**Mean**	**SE**
Low Explainability	No choice	68	18.31	<.0001	4.825	0.4146
	Choice	68	20.68	<.0001	3.3548	0.1964
High Explainability	No choice	68	12.13	<.0001	3.0357	0.278
	Choice	68	15.92	<.0001	3.4559	0.2691

Smallest Cohen's d (Comparison from Low Explainability – No-choice) = 3.88

$r_{Y\lambda} = 0.89$

Appendix 17: Study Set 2, Experiment 2, Actual AI Reliance Compared to Self-Assessed AI Reliance

Actual Reliance on the AI Recommendation

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Explainability Level	1	68	0.08	0.7777
Choice	1	68	0.51	0.4774
Explainability*Choice	1	68	0.68	0.4121

Explainability Level*Choice Least Squares Means						
Explainability Level	**Choice**	**DF**	**t Value**	**Pr > |t|**	**Mean**	**SE**
High Explainability	Choice	68	12.65	<.0001	3.5588	0.3571
	No choice	68	11.31	<.0001	3.5179	0.3912
Low Explainability	Choice	68	15.87	<.0001	3.3629	0.257
	No choice	68	11.06	<.0001	3.95	0.4905

Self-Assessed AI Reliance on the AI Recommendation

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Explainability Level	1	67	11.24	0.0013
Choice	1	67	33.86	<.0001
Explainability*Choice	1	67	0.34	0.5609

Condition Least Squares Means					
Explainability Level	**Estimate**	**Standard**	**DF**	**t Value**	**Pr > |t|**
Low Explainability	3.8981	0.1238	67	31.49	<.0001
High Explainability	3.2966	0.1299	67	25.39	<.0001

Cohen's d = 4.80

$r_{Y\lambda} = 0.92$

Choice Least Squares Means					
Choice	**Estimate**	**Standard**	**DF**	**t Value**	**Pr > |t|**
Choice	4.1193	0.1035	67	39.8	<.0001
No choice	3.0754	0.1465	67	20.99	<.0001

Cohen's d = 8.16

$r_{Y\lambda} = 0.97$

Explainability Level*Choice Least Squares Means						
Explainability Level	**Choice**	**DF**	**t Value**	**Pr > |t|**	**Mean**	**SE**
High Explainability	Choice	4.3676	0.1664	67	26.26	<.0001
	No choice	3.4286	0.1833	67	18.7	<.0001
Low Explainability	Choice	3.871	0.1232	67	31.42	<.0001
	No choice	2.7222	0.2286	67	11.91	<.0001

Smallest Cohen's d (Comparison from High Explainability – Choice) = 1.53

$r_{Y\lambda} = 0.61$

Appendix 18: Study Set 2, Experiments 1-2, Task Engagement

Panel 1: Experiment 1 Task Engagement

	Assigned	Completed
Low Explainability	37	25 (67.57%)
High Explainability	38	22 (57.89%)
Totals	**75**	**47**

$\chi^2(0.75, n = 75) = 0.39$

Panel 2: Experiment 2 Task Engagement

		Completed by Explainability Level		
	Assigned	High	Low	Total
Choice	62	17	38	55 (88.71%)
No choice	62	14	14	28 (45.16%)
Totals	**124**	**31**	**52**	**83**

$\chi^2(26.56, n = 124) < 0.001$

Endnotes

1. Office of the Director of National Intelligence, *National Intelligence Strategy of the United States of America 2019*, January 2019, https://www.dni.gov/files/ODNI/documents/National_Intelligence_Strategy_2019.pdf.
2. Amir Husain, *The Sentient Machine: The Coming Age of Artificial Intelligence* (New York: Scribner, 2017).
3. Office of the Director of National Intelligence, *National Intelligence Strategy of the United States of America 2019*.
4. Office of the Director of National Intelligence, *The AIM Initiative: A Strategy for Augmenting Intelligence Using Machines*, 2019, https://www.dni.gov/index.php/newsroom/reports-publications/item/1940-the-aim-initiative-a-strategy-for-augmenting-intelligence-using-machines.
5. Paul E. Meehl, *Clinical Versus Statistical Prediction* (Minneapolis: University of Minnesota Press, 1954).
6. Paul E. Meehl, "Causes and Effects of My Disturbing Little Book," *Journal of Personality Assessment* 50, no. 3 (1986): 370-75, https://www.tandfonline.com/doi/abs/10.1207/s15327752jpa5003_6.
7. Frank Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review* 65, no. 6 (1958): 386-408, https://www.cs.cmu.edu/~epxing/Class/10715-14f/reading/Rosenblatt.perceptron.pdf.
8. Office of the Under Secretary of Defense (Comptroller)/Chief Financial Officer, *Defense Budget Overview: United States Department of Defense Fiscal Year 2020 Budget Request*, March 2019, https://comptroller.defense.gov/Portals/45/Documents/defbudget/fy2020/fy2020_Budget_Request_Overview_Book.pdf.
9. "Future in the Balance? How Countries Are Pursuing an AI Advantage," Deloitte (online article), accessed on September 30, 2020, https://www2.deloitte.com/cn/en/pages/technology-media-and-telecommunications/articles/how-countries-are-pursuing-an-ai-advantage.html.
10. Berkeley J. Dietvorst, Massey C. Simmons, and Cade Massey, "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err," *Journal of Experimental Psychology* 144, no. 1 (February 2015): 114-26, https://pubmed.ncbi.nlm.nih.gov/25401381/.
11. Jennifer M. Logg, Julia A. Minson, and Don A. Moore, "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment," *Organizational Behavior and Human Decision Processes* 151 (March 2019): 90-103, https://www.sciencedirect.com/science/article/abs/pii/S0749597818303388.
12. Matt Stroud, "Chicago PD Automated Policing Program Got This Man Shot Twice," The Verge, May 24, 2021, https://www.theverge.com/22444020/chicago-pd-predictive-policing-heat-list.
13. "List of Artificial Intelligence Films," Wikipedia, accessed on September 15, 2021, https://en.wikipedia.org/wiki/List_of_artificial_intelligence_films.
14. J. Huber, J. W. Payne, and C. Puto, "Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis," *Journal of Consumer Research* 9, no.1 (1982): 90-98, https://psycnet.apa.org/record/1982-29226-001.
15. Itamar Simonson, "Choice Based on Reasons: The Case of Attraction and Compromise Effects," *Journal of Consumer Research* 16, no. 2 (September 1989): 158-74, https://academic.oup.com/jcr/article-abstract/ 16/2/158/1800431.
16. Jean-Luc Chabert, *A History of Algorithms: From the Pebble to the Microchip* (Berlin: Springer, 1999).

17. David A. Hounshell, *From the American System to Mass Production, 1800-1932: The Development of Manufacturing Technology in the United States* (Baltimore: Johns Hopkins University Press, 1984).
18. Liza Mundy, *Code Girls: The Untold Story of the American Women Code Breakers of World War II* (New York: Hachette Books, 2017).
19. Jessica M. Eaglin, "Constructing Recidivism Risk," *Emory Law Journal* 67, no. 1 (2017): 81, https://scholarly commons.law.emory.edu/cgi/viewcontent.cgi?article=1046&context=elj.
20. Alan M. Turing, "On Computable Numbers, With an Application to the Entscheidungs Problem," Proceedings of the London Mathematical Society 42 (1936): 230-65, https://www.cambridge.org/core/journals/journal-of-symbolic-logic/article/abs/m-turing-on-computable-numbers-with-an-application-to-the-entscheidungs-problem-proceedings-of-the-london-mathematical-society-2-s-vol-42-19361937-pp-230265/4DFCA89035F7F7C5BF4DB5129B8BB09E.
21. Virginia Apgar, "A Proposal for a New Method of Evaluation of the Newborn Infant," *Current Researches in Anesthesia & Analgesia* 32, no. 4 (July 1953): 260-67, https://journals.lww.com/anesthesia-analgesia/Citation/1953/07000/A_Proposal_for_a_New_Method_of_Evaluation_of_the.6.aspx.
22. Committee on Obstetric Practice, "The Apgar Score," *American Academy of Pediatrics* 136, no. 4 (October 2015), https://publications.aap.org/pediatrics/article/136/4/819/73821/The-Apgar-Score.
23. Meehl, *Clinical Versus Statistical Prediction*.
24. Daniel Kahneman, *Thinking Fast and Slow* (New York: Farrar, Strauss, and Giroux, 2011).
25. Susan Wharton Gates, Vanessa Gail Perry, and Peter M. Zorn, "Automated Underwriting in Mortgage Lending: Good News for the Underserved?" *Housing Policy Debate* 13, no.2 (2002): 369-91, https://www.tandfonline.com/doi/abs/10.1080/10511482.2002.9521447.
26. Dan Hurley, "Can an Algorithm Tell When Kids Are in Danger?" *New York Times,* January 2, 2018, https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html.
27. Bo Cowgill, "Automating Judgement and Decisionmaking: Theory and Evidence from Resume Screening," forthcoming Columbia Business School Research Paper (2017).
28. W. M. Grove et al., "Clinical Versus Mechanical Prediction: A Meta-Analysis," *Psychological Assessment* 12, no. 1 (March 2000): 19, https://pubmed.ncbi.nlm.nih.gov/10752360/.
29. Kat Eschner, "In 1913, Henry Ford Introduced the Assembly Line: His Workers Hated It," *Smithsonian Magazine*, December 1, 2016, https://www.smithsonianmag.com/smart-news/one-hundred-and-three-years-ago-today-henry-ford-introduced-assembly-line-his-workers-hated-it-180961267/.
30. Rockwell Anyoha, "The History of Artificial Intelligence," *Science in the News* (Special Edition: Artificial Intelligence), Harvard University, September 2017, https://sitn.hms.harvard.edu/special-edition-artificial-intelligence/.
31. Turing, "On Computable Numbers."
32. Ethem Alpaydin, *Machine Learning: The New AI* (Cambridge, MA: The MIT Press, 2016)
33. Evelyn Fix and Joseph L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," *USAF School of Aviation Medicine*, February 1951, https://apps.dtic.mil/dtic/tr/fulltext/u2/ a800276.pdf.
34. Naomi S. Altman, "An Introduction to Kernel and Nearest Neighbor Nonparametric Regression," *The American Statistician* 46, no. 3 (1992): 175-85, https://ecommons.cornell.edu/bitstream/handle/1813/31637/BU-1065-MA.pdf;jsessionid=7440F65B8A0BEB4073CC4D6675618D66?sequence=1.
35. Seymour Papert, "The Summer Vision Project," MIT Project Mac Artificial Intelligence Group Vision Memo, no. 100 (July 7, 1966), https://dspace.mit.edu/bitstream/handle/1721.1/6125/AIM-100.pdf?sequence=2&isAllowed=y.
36. Dietvorst, Simmons, and Massey, "Algorithm Aversion."
37. D. Onkal et al., "The Relative Influence of Advice From Human Experts and Statistical Methods on Forecast Judgments," *Journal of Behavioral Decision Making* 22, no. 4 (October 2009): 390-409, https://www.researchgate. net/

publication/230142140_The_Relative_Influence_of_Advice_From_Human_Experts_and_Statistical_Methods_on_Forecast_Adjustments.
38. Victoria A. Shaffer et al., "Why Do Patients Derogate Physicians Who Use a Computer-based Diagnostic Support System?" *Medical Decision Making* 33, no. 1 (January 2013): 108-18, https://pubmed.ncbi.nlm.nih.gov/22820049/.
39. Logg, Minson, and Moore, "Algorithm Appreciation."
40. Benjamin von Walter, Dietmar Kremmel, and Bruno Jager, "The Impact of Lay Beliefs About AI on Adoption of Algorithmic Advice," *Marketing Letters* 33 (August 24, 2021): 143-55, https://link.springer.com/article/10.1007/s11002-021-09589-1.
41. Jennifer M. Logg, "Theory of Machine: When Do People Rely on Algorithms?" Harvard Business School Working Paper, No. 17-086, March 2017, http://nrs.harvard.edu/urn-3:HUL.InstRepos:31677474.
42. Gary A. Klein, *Sources of Power* (Cambridge: MIT Press, 1998).
43. K. Anders Ericsson and W. Kintsch, "Shortcomings of Generic Retrieval Structures With Slots of the Type That Gobet (1993) Proposed and Modeled," *British Journal of Psychology* 91 (2000): 571-88, https://pubmed.ncbi.nlm.nih.gov/11104179/.
44. J. Anders Ericcson and W. Kintsch, "Long-Term Working Memory," *Psychological Review* 102, no. 2 (April 1995): 211-45, https://pubmed.ncbi.nlm.nih.gov/7740089/.
45. J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton: Princeton University Press, 1944).
46. Amos Tversky, "Intransitivity of Preferences," *Psychological Review* 76, no. 1 (1969): 31-48, https://pages.ucsd.edu/~mckenzie/Tversky1969PsychReview.pdf.
47. Jungkeun Kim, Jongwon Park, and Gangseog Ryu, "Decoy Effects and Brands," in *NA–Advances in Consumer Research*, edited by Connie Pechmann and Linda Price (Duluth: Association for Consumer Research, 2006): 683-87.
48. Sharoni Shafir, "Intransitivity of Preferences in Honey Bees: Support for 'Comparative' Evaluation of Foraging Options," *Animal Behaviour* 48, no. 1 (July 1994): 55-67, https://www.sciencedirect.com/science/article/pii/S0003347284712115.
49. William L. Bewley et al., "Human Factors Testing in the Design of Xerox's 8010 'Star' Office Workstation," The ACM Conference on Human Factors in Computing Systems (CHI) Proceedings, December 1983, 72-77, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.125.4373&rep=rep1&type=pdf.
50. Robert Wachter, *The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age* (New York: McGraw-Hill Education, 2017).
51. Robert C. Beck and Charles I. Brooks, "Human Judgments of Stimuli Associated with Shock Onset and Termination," *Psychonomic Science* 8 (1967): 327-28, https://link.springer.com/article/10.3758/BF03331685.
52. Harry Helson, *Adaptation-Level Theory* (Oxford, Harper & Row, 1964).
53. Philip Brickman, Dan Coates, and Ronnie Janoff-Bulman, "Lottery Winners and Accident Victims: Is Happiness Relative?" *Journal of Personality and Social Psychology* 36, no. 8, (August 1978): 917-27, https://pubmed.ncbi.nlm.nih.gov/690806/.
54. Richard L. Oliver, "Measurement and Evaluation of Satisfaction Processes in Retail Settings," *Journal of Retailing* 57, no. 3 (1981): 25-48, https://www.semanticscholar.org/paper/Measurement-and-evaluation-of-satisfaction-in-Oliver/b7f81c09b9179dadec53bf7b511e556325655126.
55. U.S. Department of Defense, "Project Maven To Deploy Computer Algorithms to War Zone by Year's End," by Cheryl Pellerin, DoD News, July 21, 2017, https://www.defense.gov/Explore/News/Article/Article/1254719/project-maven-to-deploy-computer-algorithms-to-war-zone-by-years-end/.
56. Matthew Rosenberg and John Markoff, "The Pentagon's 'Terminator Conundrum': Robots That Could Kill on Their Own," October 25, 2016, https://www.nytimes.com/2016/10/26/us/pentagon-artificial-intelligence-terminator.html.

57. Amit Datta, Michael Carl Tschantz, and Anupam Datta, "Automated Experiments on Ad Privacy Settings," *Proceedings on Privacy Enhancing Technologies* 2015, no. 1 (2015): 92-112, https://petsymposium.org/popets/ 2015/popets-2015-0007.php.
58. Kate Crawford, "Think Again: Big Data: Why the Rise of Machines Isn't All It's Cracked Up To Be," *Foreign Policy*, May 10, 2013, https://foreignpolicy.com/2013/05/10/think-again-big-data/.
59. Jinhua Tian et al., "Multidimensional Face Representation in a Deep Convolutional Neural Network Reveals the Mechanism Underlying AI Racism," *Computational Neuroscience* 15 (March 10, 2021): 1-8, https://www.frontiersin.org /articles/10.3389/fncom.2021.620281/full.
60. Soenke Ziesche, "AI Ethics and Value Alignment for Nonhuman Animals," *Philosophies* 6, no. 2 (April 13, 2021): 1-12, https://www.mdpi.com/2409-9287/6/2/31.
61. Andreas Holzinger, "Interactive Machine Learning for Health Informatics: When Do We Need the Human-in-the-Loop?" *Brain Informatics* 3 (March 2, 2016): 119-31, https://link.springer.com/content/pdf/10.1007/s40708-016-0042-6.pdf.
62. Bartlomiej Grychtol et al., "Human Behavior Integration Improves Classification Rates in Real-Time BCI," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 18, No. 4 (August 9, 2010): 362-68, https://ieeexplore.ieee.org/abstract/document/5545733.
63. P. Jonathon Phillips et al., *Four Principles of Explainable Artificial Intelligence,* NISTIR 8312, National Institute of Standards and Technology, September 2021, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933399.
64. Phillips et al., *Four Principles of Explainable Artificial Intelligence*.
65. Meehl, *Clinical Versus Statistical Prediction*.
66. J. Zhu et al., "Explainable AI for Designers: A Human-Centered Perspective on Mixed-initiative Co-Creation," 2018 IEEE Conference on Computational Intelligence and Games (CIG), August 14-17, 2018, 1-8, https://ieeexplore.ieee.org/document/8490433.
67. Katharina Weitz et al., "'Do You Trust Me?': Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design," Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, July 1, 2019, 7-9, https://doi.org/10.1145/3308532.3329441.
68. Matt Turek, "Explainable Artificial Intelligence (XAI)," Defense Advanced Research Projects Agency, July 19, 2021, https://www.darpa.mil/program/explainable-artificial-intelligence.
69. Arzam Kotriwala, et al., "XAI for Operations in the Process Industry–Applications, Theses, and Research Directions," CEUR Workshop Proceedings, 2021, 1-12, http://ceur-ws.org/Vol-2846/paper26.pdf.
70. Erika Patall and Harris Cooper, "The Effects of Choice on Intrinsic Motivation and Related Outcomes: A Meta-Analysis of Research Findings," *Psychological Bulletin* 134, no. 2 (April 2008): 270-300, https://www.researchgate.net/profile/Erika-Patall/publication/5554527_The_Effects_of_Choice_on_Intrinsic_Motivation_and_Related_Outcomes_A_Meta-Analysis_of_Research_Findings/links/004635370ddf699700000000/The-Effects-of-Choice-on-Intrinsic-Motivation-and-Related-Outcomes-A-Meta-Analysis-of-Research-Findings.pdf.
71. Donald E. Powers and Randy Elliot Bennett, "Effects of Allowing Examinees To Select Questions on a Test of Divergent Thinking," *Applied Measurement in Education* 12, no. 3 (1999), https://www.tandfonline.com/doi/abs/10.1207/S15324818AME1203_3.
72. Sheena S. Iyengar and Mark R. Lepper, "Rethinking the Value of Choice: A Cultural Perspective on Intrinsic Motivation," *Journal of Personality and Social Psychology* 76, no. 3 (1999): 349-66, https://web.mit.edu/curhan/www/docs/Articles/15341_Readings/Behavioral_Decision_Theory/Iyengar_Lepper_1999_Rethinking_the_value_of_choice.pdf.
73. Anyi Ma, Yu Yang, and Krishna Savani, "'Take It or Leave It' A Choice Mindset Leads to Greater Persistence and Better Outcomes in Negotiations," *Organizational Behavior and Human Decision Processes*, 153 (2019): 1-12, http://www.yu-yang.com/papers/Ma_Yang_Savani_2019_OBHDP.pdf.

74. Amy Morin, "10 Ways To Stop Giving People Power Over You, According to a Psychotherapist," *Business Insider*, January 9, 2020, https://www.businessinsider.com/psychotherapist-10-ways-to-stop-giving-people-power-over-you-2020-1.
75. Craig R. Fox, "The Availability Heuristic in the Classroom: How Soliciting More Criticism Can Boost Your Course Ratings," *Judgment and Decision Making*, February2006, 86-90, https://www.researchgate.net/profile/Craig-Fox/publication/5140605_The_availability_heuristic_in_the_classroom_How_soliciting_more_criticism_can_boost_your_course_ratings/links/54b93abf0cf2d11571a31e6f/The-availability-heuristic-in-the-classroom-How-soliciting-more-criticism-can-boost-your-course-ratings.pdf.
76. Lee Ross and Richard E. Nisbett, *The Person and the Situation* (London: Pinter and Martin, Ltd., 1991).
77. Steven D. Levitt and John A. List, "Homo Economicus Evolves," *Science* 319, no. 5865 (February 15, 2008): 909-10, https://www.science.org/doi/10.1126/science.1153640.
78. Turek, "Explainable Artificial Intelligence (XAI)."
79. Craig Enders, "An Introduction to Maximum Likelihood Estimation," in *Applied Missing Data Analysis* (New York: The Guilford Press, 2010), 56-85.
80. Dietvorst, Simmons, and Massey, "Algorithm Aversion."
81. Andrew F. Hayes, "Further Examples of Conditional Process Analysis," in *Introduction to Mediation, Moderation, and Conditional Process Analysis*, ed. Andrew F. Hayes (New York: The Guilford Press, 2018), 431-68.
82. W. G. Chase and H. A. Simon, "The Mind's Eye in Chess," in *Visual information Processing*, ed. W. G. Chase (New York: Academic Press, 1997), 215-81.
83. G. A. Klein and R. R. Hoffman, "Seeing the Invisible: Perceptual-Cognitive Aspects of Expertise," in *Cognitive Science Foundations of Instruction*, ed. M. Rabinowitz (Mahwah, NJ: Erlbaum, 1992), 203-26.
84. P. E. Tetlock and A. Belkin, *Counterfactual Thought Experiments in World Politics* (Princeton: Princeton University Press, 1996).
85. Dietvorst, Simmons, and Massey, "Algorithm Aversion."
86. J. G. March and H. A. Simon, "Cognitive Limits on Rationality," *Organizations*, 1958, 136-71.
87. Andrew J. Elliot and Judith M. Harackiewicz, "Approach and Avoidance Achievement Goals and Intrinsic Motivation: A Mediational Analysis," *Journal of Personality and Social Psychology* 70, no. 3 (March 1996), 461-75, https://www.researchgate.net/publication/263916494_Approach_and_Avoidance_Achievement_Goals_and_Intrinsic_Motivation_A_Mediational_Analysis.
88. Eleanor Eytam, Noam Tractinsky, and Oded Lowengart, "The Paradox of Simplicity: Effects of Role on the Preference and Choice of Product Visual Simplicity Level," *International Journal of Human-Computer Studies* 105 (September 2017): 43-55, https://www.sciencedirect.com/science/article/abs/pii/S1071581917300599.
89. Qian Zhuang et al., "Translation of Fear Reflex Into Impaired Cognitive Function Mediated by Worry," *Science Bulletin*, 61 (October 25, 2016), https://link.springer.com/article/10.1007/s11434-016-1177-9.
90. Christopher Mesagno, Jack T. Harvey, and Christopher M. Janelle, "Choking Under Pressure: The Role of Fear of Negative Evaluation," *Psychology of Sport and Exercise* 13, no. 1 (January 2012): 60-68, https://www.sciencedirect.com/science/article/abs/pii/S1469029211001038.
91. Tania Lombrozo, "Simplicity and Probability in Causal Explanation," *Cognitive Psychology* 55, no. 3 (November 2007): 232-57, https://pubmed.ncbi.nlm.nih.gov/17097080/.
92. J. Feldman, "Minimization of Boolean Complexity in Human Concept Learning," *Nature* 407 (October 5, 2000): 630-33, https://www.nature.com/articles/35036586.
93. Debora Viana Thompson, Rebecca W. Hamilton, and Roland T. Rust, "Feature Fatigue: When Product Capabilities Become Too Much of a Good Thing," *Journal of Marketing Research* 42, no. 4 (2005): 431-42, https://journals.sagepub.com/doi/10.1509/jmkr.2005.42.4.431.

94. Neil Anderson et al., "Because It's Boring, Irrelevant, and I Don't Like Computers: Why High School Girls Avoid Professionally-Oriented ICT Subjects," *Computers and Education* 50, no. 4 (May 2008), 1314, https://www.sciencedirect.com/science/article/abs/pii/S0360131506001953.
95. K. Anders Ericsson and R. Pool, *Peak: Secrets From the New Science of Expertise* (Boston: Mariner Books, 2016)
96. Jeong-Yeol Park and SooCheong (Shawn) Jang, "Confused by Too Many Choices? Choice Overload in Tourism," *Tourism Management* 35 (April 2013): 1-12, https://www.sciencedirect.com/science/article/abs/pii/S0261517712000921.
97. Barbara L. Fredrickson and Daniel Kahneman, "Duration Neglect in Retrospective Evaluations of Affective Episodes," *Journal of Personality and Social Psychology* 65, no.1 (1993): 45-55, https://doi.apa.org/doiLanding?doi=10.1037%2F0022-3514.65.1.45.
98. Donald A. Redelmeier and Daniel Kahneman, "Patients' Memories of Painful Medical Treatments: Real-Time and Retrospective Evaluations of Two Minimally Invasive Procedures," *Pain* 66, no.1 (July 1996): 3-8, https://journals.lww.com/pain/Abstract/1996/07000/Patients__memories_of_painful_medical_treatments_.2.aspx.
99. Michael D. Anestis et al., "A Comparison of Retrospective Self-Report Versus Ecological Momentary Assessment Measures of Affective Lability in the Examination of Its Relationship with Bulimic Symptomatology," *Behaviour Research and Therapy* 48, no. 7 (July 2010): 607-13, https://www.sciencedirect.com/science/article/abs/pii/S000579671000046X?via%3Dihub.
100. Kevin Doherty and Gavin Doherty, "The Construal of Experience in HCI: Understanding Self-Reports," *International Journal of Human-Computer Studies* 110 (2018), 63-74, https://www.scss.tcd.ie/Gavin.Doherty/papers/Selves_IJHCS_2018.pdf.
101. Amos Tversky and Daniel Kahneman, "Belief in the Law of Small Numbers," *Psychological Bulletin*, 76 (August 1971): 105-110, https://www.semanticscholar.org/paper/BELIEF-IN-THE-LAW-OF-SMALL-NUMBERS-Tversky-Kahneman/894fc603f9b16e775f95045fb805b5d7e6935944.
102. Sheena S. Iyengar and Mark R. Lepper, "When Choice Is Demotivating: Can One Desire Too Much of a Good Thing?" *Journal of Personality and Social Psychology* 79, no. 6 (December 2000): 995-1006, https://pubmed.ncbi.nlm.nih.gov/11138768/.
103. Dietvorst, Simmons, and Massey, "Algorithm Aversion."
104. Logg, Minson, and Moore, "Algorithm Appreciation."