AI Trust and Alignment in High-Stakes Decision-Making Environments

Katelyn R. Smith
University of Massachusetts Lowell
Lowell, MA
Katelyn_Smith@uml.edu

AnaCristina Bedoya
University of Massachusetts Lowell
Lowell, MA
AnaCristina_Bedoya@uml.edu

Neil Shortland
University of Massachusetts
Lowell, MA
Neil_Shortland@uml.edu

Joseph Cohn
SoarTech
Ann Arbor, Michigan
Joseph.Cohn@soartech.com

Robert Bixler
SoarTech
Ann Arbor, Michigan
Robert.Bixler@soartech.com

Jordan Lampi
SoarTech
Ann Arbor, Michigan
Jordan.Lampi@soartech.com

Angela Woods
SoarTech
Ann Arbor, Michigan
Amgela.Woods@soartech.com

ABSTRACT

Decision-making in high-stakes environments depends on effectively managing cognitive load and minimizing uncertainty. Artificial Intelligence (AI)-enabled decision-support tools are increasingly used in military and civilian settings to assist with complex tasks, but their success depends on trust and alignment between users and AI systems. In this pilot study, 263 civilian participants with no prior medical or triage training, intentionally chosen to examine trust formation without domain expertise, completed simulated medical triage scenarios. Participants were randomly assigned to one of three conditions: (1) explanations for AI recommendations hidden behind a button; (2) explanations displayed by default; or (3) no explanations provided. Each participant evaluated AI triage decisions that either aligned or misaligned with their own choices.

Ordinal logistic regression indicated that alignment significantly increased trust (OR = 1.87, p = .02), and absence of explanation was associated with lower trust (OR = 0.64, p = .10). Explanations had the greatest effect when participants were already aligned with the AI, suggesting they reinforce rather than create trust. These findings informed a proposed framework for human-AI trust in high-uncertainty, high-stakes contexts, emphasizing (1) behavioral alignment, (2) contextually relevant explanations, and (3) dynamic adaptation to user decision styles. This framework extends existing models by focusing on morally complex, time-pressured environments where traditional rational-choice assumptions fail. We discuss implications for designing AI systems that maintain calibrated trust and effective integration into critical decision workflows.

PUBLIC RELEASE STATEMENT

The research reported in this document/presentation was performed in connection with contract number W912CG- 23-C-0036 with the U.S. Army Contracting Command - Aberdeen Proving Ground (ACC-APG) and the Defense Advanced Research Projects Agency (DARPA). The views and conclusions contained in this document/presentation are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of ACC-APG, DARPA, or the U.S. Government unless so designated by other authorized documents.

ABOUT THE AUTHORS

Katelyn R. Smith is a PhD candidate at UMass Lowell's Center for Terrorism and Security Studies. Her research focuses on irregular warfare, security force training, and hybrid conflicts. She manages a dataset of 10,000+ extremist acts in the U.S. and develops counterterrorism training for military and law enforcement. Katelyn regularly presents internationally and collaborates across disciplines to connect research with practice.

AnaCristina Bedoya is a PhD candidate in criminology at the University of Massachusetts Lowell. She has been funded as a research assistant on several grants from the United States Army and Department of Defense, providing expertise in quantitative methods, psychophysiological methods, and survey data collection. AnaCristina is a behavioral scientist and victimologist who is especially interested in studying the coping behaviors of survivors of sexual assault. Her most recent research project is a mixed methods exploration of sexual scripts of survivors of sexual assault.

Neil Shortland, PhD, is Director of the Center for Terrorism and Security Studies at UMass Lowell and an Associate Professor of Criminology. With experience in UK defense and policing, his research focuses on decision-making in security contexts. He co-developed the LUCIFER tool to study individual differences in high-stakes decisions, supported by the U.S. Army and NSF, and has applied it to emergency triage with DARPA.

Joseph Cohn, PhD, is Director of SoarTech's Readiness and Medical Solutions team and is a retired Navy Medical Service Corps Captain. He specializes in biomedical and human-machine interface solutions using AI, brain-machine interfaces, and wearable sensors to support military and medical applications. Joseph is an Associate Fellow of the Aerospace Medical Association and a Fellow of both the Society for Military Psychology and the American Psychological Association.

Robert Bixler is a Research Scientist at SoarTech with experience in machine learning (ML), cognitive agents, adversarial machine learning, and application of AI/ML to cybersecurity. He has researched attacks and defenses against ML models, game theory/reinforcement learning approaches to learn cyberspace courses of action, and techniques to understand social media discourses. He has published papers in peer-reviewed conferences and journals in the fields of human-computer interaction, intelligent tutoring systems, and cybersecurity, including papers on developing computational models of cyberspace-based tactics, techniques, and procedures.

Jordan Lampi is a software engineer with 13+ years of experience developing autonomous systems that support humans in complex, data-limited environments. He specializes in multi-agent coordination, autonomy frameworks, and HMIs, with a focus on aligning AI behavior with human workflows. He has led development and integration on high-impact programs and managed interdisciplinary teams.

Angela J. Woods is a Software Architect at Soar Technology with 25+ years of experience in real-time systems, intelligent training, simulation, and data analytics. She specializes in adaptive architectures and machine learning for causal modeling and visualization. Angela has led over $20M in engineering efforts and architected solutions on $50M+ in government-funded research.

AI Trust and Alignment in High-Stakes Decision-Making Environments

Katelyn R. Smith
University of Massachusetts Lowell
Lowell, MA
Katelyn_Smith@uml.edu

AnaCristina Bedoya
University of Massachusetts Lowell
Lowell, MA
AnaCristina_Bedoya@uml.edu

Neil Shortland
University of Massachusetts
Lowell, MA
Neil_Shortland@uml.edu

Joseph Cohn
SoarTech
Ann Arbor, MI
Joseph.Cohn@soartech.com

Robert Bixler
SoarTech
Ann Arbor, MI
Robert.Bixler@soartech.com

Jordan Lampi
SoarTech
Ann Arbor, MI
Jordan.Lampi@soartech.com

Angela Woods
SoarTech
Ann Arbor, MI
Amgela.Woods@soartech.com

INTRODUCTION

Effective decision-making in high-stakes environments– such as mass casualty events, military operations, and crisis response– requires individuals to make rapid, morally complex judgments under high cognitive load and persistent uncertainty (Johnson, 2020; Dodig Crnkovic et al., 2025). Artificial intelligence (AI)-enabled decision-support tools are increasingly deployed to help reduce this burden and support decision quality in these environments (Dingel et al., 2024; Pedro et al., 2023). These systems promise improved outcomes by aiding complex, time-sensitive judgments, but their success depends on more than just performance metrics. Trust and alignment between human users and AI systems have emerged as critical factors influencing system uptake and effective integration (Koenig, 2025; Stevens & Stetson, 2023). Trust refers not only to a user's willingness to rely on AI, but also to their perception of the AI's competence, transparency, and shared reasoning. Alignment– whether users perceive that the AI makes similar decisions or shares their goals– plays a key role in calibrating trust appropriately (Corvelo Benz & Gomez Rodriguez, 2024; Song & Lin, 2023). These dynamics are especially important across different human-AI teaming architectures, including in-the-loop, on-the-loop, and off-the-loop decision-making (Fuchs et al., 2025). In conditions of uncertainty, trust miscalibration– either over- or under-reliance– can have serious consequences (Zhang et al., 2020; Spitzer et al., 2025).

This study investigates how AI-generated explanations and perceived decision alignment influence user trust in a simulated medical triage context. It was conducted as part of the Defense Advanced Research Projects Agency's (DARPA) *In the Moment (ITM)* program (McVay, 2025), which investigates whether aligning AI systems to individual human decision-makers increases willingness to delegate in high-stakes domains such as medicine and national security contexts where experts often disagree and decisions must be made under ambiguity and pressure.
Using medical triage decision scenarios, we evaluate whether transparent explanations enhance trust, particularly when synthetic decision-makers make the same choices as participants. We hypothesized that adding explanations for these decisions will enhance human trust. Our findings contribute empirical insight and introduce a framework for designing AI systems that align more effectively with human reasoning, improve trust calibration, and support responsible AI integration in high-stakes decision workflows (Fügener et al., 2021; Hemmer et al., 2023).

SYSTEM OVERVIEW & BACKGROUND

The 2023 National Artificial Intelligence (AI) Research and Development Strategic Plan states that "AI is the most powerful tool of our time" (p. vii). Similarly, the 2021 National Security Commission on AI explains it as "an inspiring technology. It will be the most powerful tool in generations for benefiting humanity…. 'game changing' is not a cliché." In every domain, practitioners, scholars, and politicians are grappling with what a future AI-enabled world will look like. Nowhere is the integration of AI more prevalent than within the study of human decision-making. Based on the widespread knowledge of inadequacies in the human decision-making process (e.g., Tversky & Kahneman, 1974), developers, psychologists, economists, ethicists, and computer scientists (among many others) have grappled with how we can improve human decision-making through the integration of AI, or replacement of humans with AI.

In this rapidly evolving field of study, two significant changes have created a pressing need for new psychological research on AI and decision-making. First, the role of the human is being diluted within the decision-making cycle with a vision to increasingly autonomize AI, in which the human is off the loop. Secondly, the type of decision that AI is being used to tackle is becoming increasingly complex. The recent "In the Moment" (ITM, DARPA, 2025) research program by the Defense Advanced Research Agency (DARPA) serves as one example. The ITM program seeks "technology development that supports the building, evaluating, and fielding of algorithmic decision-makers that can assume human-off-the-loop decision-making responsibilities in difficult domains, such as combat medical triage." Here, difficult domains are defined as "those where trusted decision-makers disagree; no right answer exists; and uncertainty, time-pressure, resource limitations, and conflicting values create significant decision-making challenges." A fundamental aspect of the program, and indeed the future of AI, is that when facing complex problems, in which "right" and "wrong" is perhaps subjective, it is critical that AI is aligned to the decisions that a specific human would make (noting that this decision could be different between decision-makers).

Programs such as ITM speak to the future of AI and the importance of embracing human variation within decision-making. In response to this future, this project has three goals. First, it will identify how individual differences impact decision-making in difficult domains. Second, it will identify the degree to which individual differences impact willingness to delegate to AI during difficult decision-making. Third, this project will test if individuals are more likely to delegate to an AI when it is aligned to them and mimics the effect of these individual differences. Overall, this project will support ongoing research on decision-making, trust, AI, explainable AI, and decision-making, as well as applied issues of technology development and policy on AI development. In addition, through a series of broader impact activities, this project will support the involvement of underrepresented groups in AI, while also challenging current barriers to trust with AI at a societal level.

Difficult Decision-Making

On October 1st, 2017, Stephen Paddock opened fire onto a crowd of approximately 22,000 concertgoers, killing 60 people and injuring 867 (411 of whom were wounded by gunfire). Of these 867 injured civilians, the nearby Sunrise Hospital received more than 200 penetrating gunshot wound victims. Over the next several hours, a small team of emergency physicians, trauma surgeons, residents, registered nurses, and nurses used minimal resources to make a series of rapid triage and mass casualty management decisions to save as many lives as possible. This situation typifies the challenges that decision-makers face in real-world difficult domains in which the decision-maker has high uncertainty, low experience of that exact problem, no good policy or guidance, and a need to innovate and adapt their decision-making within a rare and variable environment. In addition, such decision-makers often face multiple unappealing options, none of which guarantee a "ideal" or even "satisfactory" outcome.

Looking at these types of difficult decisions, two assertions are true: (1) there is an urgent need for autonomous, off-the-loop decision-making AI in contexts where decisions must be made within seconds and human oversight would introduce unacceptable delays, but where alignment with trusted decision-makers remains essential to reduce cognitive burden and minimize error; and (2) modelling the process through which a human makes such decisions (and to which the AI must be aligned) stretches our current conceptualization of AI development. This is because decision-making in difficult situations violates many of the pre-requisite assumptions of both "rational" and "recognition"-based models of decision-making (Shortland et al., 2018; Shortland et al., 2019; Alison & Shortland, 2021).

Instead, effective decision-making in these domains requires an individual to work through a series of sequential cognitive processes, each of which creates the opportunity for unique human variance between decision-makers. Using previous naturalistic research, we can frame these processes through the SAFE-T model (i.e., decision making emphasizing naturalistic decision making and its potential for cross-comparative analysis of incidents– van Der Heuvel et al., 2012; Alison et al., 2018). The model proposes that optimal decision-making involves four key phases: Situation Assessment (SA), Plan Formulation (PF) and Plan Execution (PE), followed by an incremental and transitional team learning (T). Within this model, a range of individual, situational and environmental factors impact the way in which an individual navigates these stages and the decision they eventually make.

Our own research using qualitative (Waring et al., 2013; Shortland et al., 2019; Shortland & Alison, 2020) and quantitative approaches (Shortland et al., 2020a, 2020b, 2021) has shown that within each of these processes, expert decision-makers will deviate significantly on how they engage with information and assess the situation, the choice they ultimately select, and when they decide to implement the decision (Power & Alison, 2017; Shortland et al., 2019). This places new challenges on those involved with creating decision-support systems which can no longer focus on "rules" and "right" answers. Instead, they must move to more dynamic decision-making strategies that reflect how experts navigate uncertainty.

Artificial Intelligence

To date, AI is present in nearly every aspect of modern society. AI affects all industries and economies, and with the exponential growth in computing infrastructure and the ever-increasing availability of "big data," the complexity, capability, and sophistication of AI is increasing in new and profound ways (Weinberger, 2019). Significant efforts have been made to enhance, fuse, and even replace human-decision making with AI. In healthcare, AI algorithms use predictive analysis algorithms to filter, organize, and search for patterns in big data sets providing a probability analysis to support decision-making (Lysaght et al., 2019). AI is already seen as a driver for the transformation of decision-making and information-centric processes (Kelly, 2012; MacCrory et al., 2014). It is already argued that "for any given skill one can think of, some computer scientist may already be trying to develop an algorithm to do it" (MacCrory, Westerman, Alhammadi, & Brynjolfsson, 2014, p. 14). Thought leaders suggest AI will take over most human jobs in the near-term future (Leetaru, 2016).

Within the explosion of AI development and adoption, one area that has garnered significant attention is the integration of AI into human decision-making. From an AI perspective, it is widely argued that AI can outperform humans in many of these domains because it is not subject to the same biases, heuristics, emotions and cognitive limitations as humans (Alufaisan et al., 2021). While we remain resistant to fully autonomous "off the loop" AI decision-making, which involves AI making decisions without any human oversight (Crootof et al., 2023), a significant focus of research has been to understand how, where, and when humans and AI can work together to make decisions.

But we are increasingly seeing evidence that autonomous decision-making AI is already being developed and deployed. For example, a United Nations Security Council report claims the Kargu-2 was used in Libya to mount autonomous attacks on human targets (Gevorgyan, 2023). These new-wave AI-enabled drones were able to "attack targets without requiring data connectivity between the operator and the munition." In this sense the decision to kill was made independently from the human operator. Such technologies, both present in war and increasingly being developed elsewhere (Abràmoff et al., 2020), present a quantitative shift in the relationship between the AI and the human from "in the loop" to an autonomous AI operating independently from human oversight.

Delegation

A central part of the human-AI team is the willingness to delegate tasks to an AI. Delegation involves transferring authority to an external agent, and therefore losing a degree of control (Leana, 1986; Hales, 1999; Richardson et al., 2002). As humans, we tend to be reluctant to give up control (Steffel et al., 2016), and hence a pervasive research challenge is understanding when and why humans will (and will not) delegate to an AI. It is largely held that a human must trust an AI to delegate a task (Glikson & Woolley, 2020). This perspective derives from core elements of the updated perspectives on the Technology Acceptance Model (TAM; Davis, 1989) in which usefulness, ease of use, and trust (Ghazizadeh, Lee, & Boyle, 2012; Hoff & Bashir, 2015) determines a users' attitudes and behavioral intentions to use a technological system (Davis, 1989). Accordingly, research has explored how physical appearance (Bainbridge et al., 2011; Lee et al., 2006), reliability (Salem et al., 2015) and personalization (Fenster et al., 2012) impact trust in an AI system. Trust is especially important, given that individuals cannot rely on "classic" personal characteristics such as goal congruence (Yukl & Fu, 1999). Instead, individuals face a "black box" that operates in a way that most, including many data scientists, cannot understand (Davenport, 2016). This is why many have moved towards the importance of "explainable AI" which explains and justifies the rationale of its predictions and actions.

Decision delegation– in theory– is assumed to be a rational cost benefit calculation. However, research is increasingly showing such decisions to be far from rational, and often suboptimal (Bobadilla-Suarez et al., 2017; Dominguez-Martinez et al., 2014). For example, Bigman and Gray (2018) found people prefer humans over AI technology to make decisions regarding life and death in the domains of driving, law, medicine, and the military– a finding partially explained by the lack of agency and emotional experience machines have. Such reduced abilities of the mind prohibit the capabilities of AI machines to engage in moral thought and thus inhibit their capacity to make moral decisions. Conversely, this interpretation can also be applied to the other side of the argument. Candrian and Scherer (2022) found it was, in fact, the perceived lower intentional capacities of AI machines (i.e., being desolate of feelings, motivation, and emotions), that meant human delegators were more likely to delegate and entrust in an AI over a human agent to decision make.

Lower capacities imply greater controllability of an AI agent reassuring the delegator the AI agent will less likely undermine them and/or act out of line or within their own self-interests making delegation less “risky.” Critically, trust then becomes a key component within delegation due to the transfer of power from the delegator to the entrusted agent. As such, perceptions regarding the agent’s capabilities, as well as the alignment of interests between the two subjects, are imperative (Lupia, 2015). Which interpretation one follows is driven by individual preferences and traits such as preference for control (Owens et al., 2014), fear of betrayal (Bulter & Miller, 2018), one’s subjective beliefs and cognitions (e.g., the tendency to trust one’s own judgment and intuition over another’s; Yaniv & Klienberger, 2000), as well as overconfidence, preference for control, and loss aversion.

These findings have important implications for understanding when and why people will prefer to delegate to an AI rather than a human counterpart. However, they also show that there are countering findings in the field. For example, Bigman and Gray (2018) found people prefer humans over AI technology, whereas Candrian and Scherer (2022) found a preference for AI. It is clear that the decision to delegate is multi-faceted and involves interpretative processes such as subjective assessments of control, fear of loss, and degree of social risk. As such, it makes sense that factors at the individual or environmental level impact these judgments and will then impact peoples’ preferences to delegate to humans vs. AI agents. However, as outlined by Candrian and Scherer (2022), research examining the effect of different types of agents (in particular, AI vs. human agents) is almost non-existent in this literature (Belanche et al., 2020), makning research examining the effect of individual differences on this process even more scant.

Explanations

In high-stakes environments where decisions carry moral weight, uncertainty, and time pressure, trust in AI systems is essential to successful human-AI teaming. One strategy for cultivating that trust is the use of explanations: descriptions of how and why an AI system generated a given recommendation. Explanations are often assumed to facilitate trust by making AI systems more transparent and comprehensive. However, growing empirical evidence suggests the relationship between explanation and trust is complex and context dependent.

Explanations may increase initial trust by enhancing the perceived transparency and competence of the system. When users are given insight into the AI’s reasoning– particularly in domains like medical triage or combat decision-making– they may feel more confident delegating critical decisions (Dingel et al., 2024; Stevens & Stetson, 2023). In our study, we explored whether such explanations also improve calibrated trust, or when users trust the AI more when appropriate and resist over-reliance when AI errors are likely. Existing research suggests that explanations modestly enhance trust, especially when the AI’s recommendation aligns with the user’s own judgment (Zhang et al., 2020; Corvelo Benz & Gomez Rodriguez, 2024). In other words, alignment and explanation work together to foster a sense of shared reasoning and legitimacy.

But explanation is not a guaranteed solution. As Zhang et al. (2020) found, even when AI systems present their confidence or rationale, users may lack the expertise or context to evaluate it meaningfully– particularly if they have no experience in the decision-domain. In our study, as in Zhang et al. (2020), participants were civilians with no triage background. In such cases, explanations that are too technical or too vague may be dismissed, misunderstood, or lead to over-trust: a false sense of security in the AI’s output. Explanations may support case-specific trust calibration, not just model-wide confidence.

Beyond clarity, user experience and training also matter. Mahmood et al. (2021) found users who participated in mock model training– where they interacted with AI training data and learned how the model operated– developed more accurate mental models and expressed greater comfort and trust in AI systems. This supports the idea that explanations are most effective when paired with user-centered design, hands-on exposure, and system feedback loops. Explanations may also affect trust differently depending on user psychology. Some may find AI's apparent rationality reassuring, while others prefer human agents for decisions involving life, death, or moral ambiguity (Bigman & Gray, 2018). Others may prefer AI precisely because of its lack of emotional bias (Candrian & Scherer, 2022). These tensions reinforce the need to tailor expectations not only to the task, but also to the user– offering layered, adjustable explanations that support varied cognitive styles and risk tolerances.

Because explanations can enhance trust and improve delegation decisions, their impact depends on their content, delivery, context, and user characteristics. They may be most powerful when combined with alignment and structured to support trust calibration, not blind reliance.

This Study

Based on the literature outlined above, this study investigates how AI-generated explanations and perceived decision alignment influence user trust in a simulated medical triage context. Using paired decision scenarios, we evaluate whether transparent explanations enhance trust, particularly when synthetic decision-makers make the same choices as participants. We test the following hypotheses:

Hypothesis 1. The presence of explanations will increase user trust in the AI decision-maker compared to conditions in which no explanation is provided.

Hypothesis 2. Alignment will mitigate the need for explanations, meaning that when people are *aligned* they will have a higher baseline trust than those who are unaligned.

Hypothesis 3. Alignment and explanations will be additive, meaning that individuals will have the most trust in AI when they are aligned and they explain their behavior.

EXPERIMENT & METHODS

Participants

Participants are a sample of 263 individuals recruited from Amazon MTurk (45 participants) and Reddit (218 participants). Participants have no previous medical or triage training and experience. Cases with incomplete data or quality issues were removed. The study was implemented on Qualtrics. Portals such as MTurk contain groups of people who are requested to participate in the completion of specific tasks, often taking the form of studies through the internet using computer mediated software (Baker et al., 2013). These pools of recruits are typically incentivized to complete tasks or to participate in studies. Literature suggests that when adhering to the best practices for MTurk pools, concerns about attention of study participants are not typically warranted (Hauser & Schwarz, 2015).

Procedure

All participants performed text triage scenarios (4 Scenarios, 1-4a; see Figure 1) and a trust survey to indicate their level of trust in a synthetic decision maker. We presented similar triage scenarios to human and synthetic decision maker pairs (i.e., Scenarios 1a and 1b, 2a and 2b, 3a and 3b, 4a and 4b). Across pairs, the synthetic decision maker made the same or opposite decision. Each participant saw the scenarios in the same order. In the trust survey, half of them were presented with delegation material where the first option was selected, and the other half were first presented with a delegation material where the second option was selected. In each case, we used the same trust survey wording and structure to determine the participants' trust level for the synthetic decision maker.

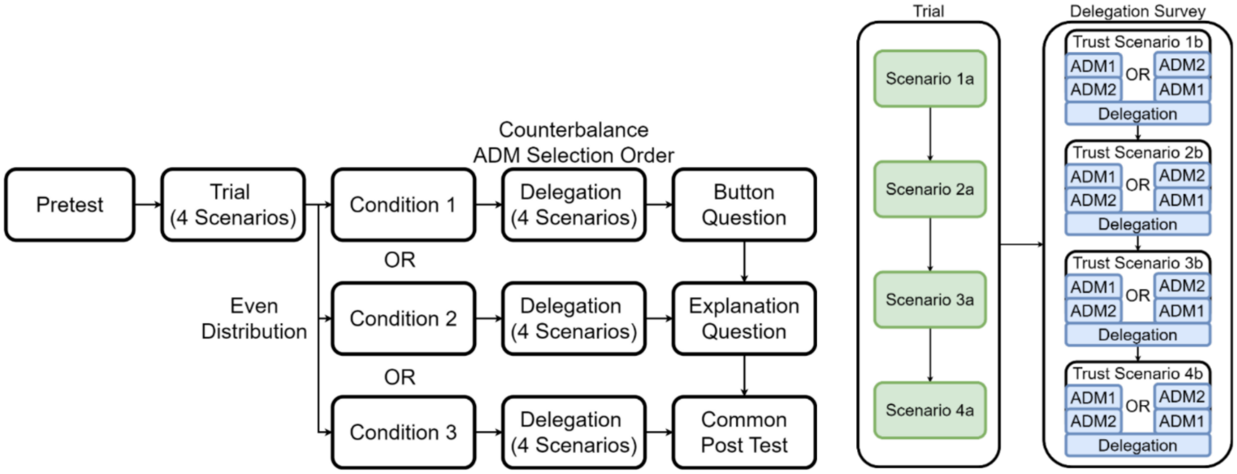

Figure 1. Experiment Flow & Scenario Ordering

In the trust survey, participants reviewed decisions made by hypothetical (synthetic) decision-makers. They were presented with scenarios where each synthetic decision-maker made the same or opposite decision with the participant. We tested three conditions that varied as a function of added explanations for the choice selected by the synthetic decision maker: In Condition 1, explanations for the choice selected by the synthetic decision maker were hidden behind a button. In Condition 2 explanations were displayed by default. In Condition 3, there were no explanations displayed.

Decision-Making Task

We designed 4 pairs of triage scenarios, 1-4a and 1-4b. In one example, scenario 1a would be shown to the participant. Then in the trust survey, the participant would be shown scenario 1b. The casualties are similar across scenarios, but the specifics are different. In scenario 1a, the casualty in the first option is a child who would have a better quality of life if treated. Similarly, the casualty in the first option for scenario 1b is also a child who would have an improved quality of life if treated, but the actual injury, cause of the injury, and the child’s gender, are all different. This was done so participants did not immediately recognize that the nature of the choice was similar to a scenario they had just seen.

Alignment Score & Trust Rating

Alignment score quantified how well the selected choices (i.e., responses) of an individual decision maker relate to the selected choices for the synthetic decision maker in the pair. An alignment score of 1 was assigned if the human and the synthetic decision maker made the same decision, and an alignment score of 0 was assigned if they made an opposite decision. Trust ratings were captured on a 5-point Likert scale, where participants could select from the following options: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree. Each response captured participants' levels of agreement with statements about the synthetic decision-maker's trustworthiness and alignment with their decision-making approach.

RESULTS

Table 1 shows the correlation between trust rating and alignment score for each condition. We found that alignment score and trust are correlated for both Reddit data and the MTurk data (conditions 2 and 3). In general, these results are promising because they demonstrate a significant correlation between alignment and trust.

Table 1. Correlation Between Trust & Alignment Score

MTurk Sample			*Reddit Sample*		
Condition	**r**	**p**	**Condition**	**r**	**p**
1 (n = 12)	.103	.319	**1** (n = 73)	.166	<.001***
2 (n = 14)	.248	.008**	**2** (n = 67)	.177	<.001***
3 (n = 19)	.266	0.001***	**3** (n = 78)	.122	.002**
Combined (n = 45)	.220	<.001***	**Combined** (n = 218)	.153	<.001***

Table 2 below shows the results of a random effects multilevel model ordinal logistic regression. Random effects were estimated for participants and scenarios to account for the same participants viewing the same scenarios, as it is assumed that the same participants will respond similarly across all the scenarios and the same scenarios may lead to similar responses across different participants. Ordinal logistic regression assumes that the change from one outcome to another (e.g., from 0 to 0.25 or 0.75 to 1) is the same. In this case, the results indicate that those who were assigned an alignment score of 1 are 1.87 times as likely to have higher trust scores. Those who were in condition 3, having no explanation provided, were 0.64 times less likely to have a higher trust score.

Table 2. Multilevel Ordinal Logistic Regression: Explanations on Trust

	Estimate	**SE**	**OR**	**p**
Alignment Score 1	0.63	0.28	1.87	0.02*
Condition 2 (explanation)	-0.06	0.28	0.94	0.82
Condition 3 (no explanation)	-0.45	0.27	0.64	0.10
Alignment Score 1 * Condition 2	0.05	0.38	1.05	0.89
Alignment Score 1 * Condition 3	0.34	0.37	1.41	0.36

DISCUSSION & RECOMMENDATIONS

In the rapidly evolving field of AI and decision-making, two significant changes have recently occurred which create a pressing need for new psychological research on *how* and *when* people will delegate to AI. It is clear that the roles we give AI and the situations in which we deploy AI will get more complex and come with higher stakes. As such, it remains paramount that the models we develop of the human-AI relationship become as dynamic and accurate as possible. This study simply presented one such extension to the field in focusing on the role of explanations and alignment, but the results here imply the need for much more research in these domains. Furthermore, we also found support for our hypothesis. Specifically:

Hypothesis 1. The presence of explanations will increase user trust in the AI decision-maker compared to conditions in which no explanation is provided. *In condition three individuals had the lowest levels of trust.*

Hypothesis 2. Alignment will mitigate the need for explanations, meaning that when people are *aligned,* they will have a higher baseline trust than those who are unaligned. *Alignment increased trust in both samples and across multiple conditions.*

Hypothesis 3. Alignment and explanations will be additive, meaning that individuals will have the most trust in AI when they are aligned and they explain their behavior. *There was no clear support for this hypothesis.*

Overall, we found general support for the importance of both alignment and explanations in that individuals rated the agent they observed with higher trust when they were aligned. Furthermore, in addition to alignment, it was also important that individuals were exposed to explanations of the AI that they were observing. These findings contribute to the much larger discussion of explainable AI (XAI). XAI refers to methods and techniques in the application of AI such that the results of the solution can be understood and interpreted by humans. As AI systems become more complex and are increasingly deployed in high-stakes domains like healthcare, finance, criminal justice, and national security, the need for transparency and trust has grown. XAI seeks to address the "black box" nature of many AI models by providing insights into how decisions are made, why certain outputs are produced, and what factors influence the outcomes. What is important here is that in the face of increasing calls for alignment, we should not view this as a substitution for XAI, but instead note that both may be required for trust when working alongside AI agents, especially in situations in which the stakes are high and the decisions are difficult.

While our findings support the notion that explanations and alignment jointly contribute to AI trust, they also highlight the nuanced nature of this relationship. Specifically, explanations appeared to have the greatest impact when participants were already behaviorally aligned with the AI's recommendation. This suggests that explanations may reinforce an existing sense of shared reasoning rather than creating trust where none exists. In contrast, for users who disagreed with the AI's decision, explanations did not significantly increase trust, indicating that misalignment may override the benefits of transparency. These findings point to a ceiling effect in the role of explanation: its influence may depend less on the explanation's presence and more on its perceived validity, relevance, and match with the user's values or decision style. This reinforces the importance of personalized, user-sensitive AI systems that not only explain themselves but also understand and adapt to the decision-maker's frame of reference.

CONCLUSION

It is clear that further research is needed in this area. This pilot study involving a mock triage task, in which alignment was conceptualized as aligned behavior, meaning "did the agent do as I do." It is important that as we increase the scope of what we are aligning to (e.g., Shortland et al., 2025), we continue to test the role of explanation and where it is (and indeed is not) additive. It remains important to look at the context of decisions and how the role of trust, alignment and explanations may vary as stakes change and decisions move from immediate decisions such as triage, to more strategic decisions associated with managing a triage event. Overall, this study shows the importance of thinking both about the alignment between the individual and the agent, and the communication that must occur as they work dynamically together.

As AI becomes more deeply embedded in complex decision-making environments, especially those marked by moral ambiguity and time pressure, fostering trust through alignment and explanation will remain a critical challenge. This study offers preliminary evidence that users trust AI systems more when they both understand and agree with them. However, unlike much of the existing literature, which has examined explanations and alignment largely in controlled or low-stakes settings, our work focuses on morally complex, high-stakes, and uncertainty-laden decisions where conventional trust-building strategies may not translate directly. Our findings suggest that explanation alone is insufficient; it must be contextually relevant and coupled with behavioral alignment to meaningfully enhance trust. Future work must explore how alignment can be dynamically modeled and operationalized in real-time systems, and how explanations can be optimized for diverse users facing high-stakes decisions. Ultimately, designing human-AI systems that are trusted, aligned, and explainable is not just a technical challenge but a human one, requiring sensitivity to human psychology, ethics, and situational complexity.

Future research should prioritize three next-phase studies. First, field experiments in real or simulated high-pressure environments (e.g., military exercises, emergency medicine drills) to test whether the alignment–explanation relationship holds under time stress and resource scarcity. Second, longitudinal studies to examine how trust evolves as users repeatedly interact with an aligned AI over time, including after AI errors. Third, development and testing of adaptive explanation systems that tailor content and depth in real-time based on user expertise, decision style, and current alignment level. These concrete steps will help translate the present findings into deployable systems that maintain trust and effectiveness when it matters most.

Ultimately, designing human-AI systems that are trusted, aligned, and explainable is not just a technical challenge but a human one, requiring sensitivity to human psychology, ethics, and situational complexity.

REFERENCES

Abràmoff, M. D., Tobey, D., & Char, D. S. (2020). Lessons learned about autonomous AI: finding a safe, efficacious, and ethical path through the development process. *American Journal of Ophthalmology, 214*, 134-142. https://doi.org/10.1016/j.ajo.2020.02.022

Alison, L. J., Palasinski, M., Humphrey, A., Humann, M., Shortland, N. D., & Bowman Grieve, L. (2018). Between a rock and a hard place of geopolitically sensitive threats – critical incidents and decision inertia. *Behavioral Sciences of Terrorism and Political Aggression, 10*(3), 207 – 224. https://doi.org/10.1080/19434472.2017.1373690

Alison, L. J., & Shortland, N. D. (2021). *Decision Time: How to make the decisions your life depends on.* Penguin (Vermillion): London, UK.

Alufaisan, A., Alrajeh, A., Alharthi, A., Alsubaie, B., Almugren, A., & Mokbel, M. F. (2021). Don't explain without verifying veracity: An analysis of explainable AI (XAI) explanations. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(8), 7166–7174. https://doi.org/10.1609/aaai.v35i8.16819

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics, 3*(1), 41–52. https://doi.org/10.1007/s12369-010-0082-7

Belanche, D., Casaló, L. V., Flavián, C., & Schepers, J. (2020). Service robot implementation: A theoretical framework and research agenda. *The Service Industries Journal, 40*(3–4), 203–225. https://doi.org/10.1080/02642069.2019.1672666

Bigman, Y. E., Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21-34. https://doi.org/10.1016/j.cognition.2018.08.003

Bobadilla-Suarez, S., Sunstein, C. R., & Sharot, T. (2017). The intrinsic value of choice: The propensity to under-delegate in the face of potential gains and losses. *Journal of risk and uncertainty, 54*(3), 187–202. https://doi.org/10.1007/s11166-017-9259-x

Butler, J. V., & Miller, J. B. (2018). Social risk and the dimensionality of intentions. *Management Science, 64*(6), 2787-2796. https://doi.org/10.287/mnsc.2016.2694

Candrian C., & Scherer, A. (2022). Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior, 134*, 107308. https://doi.org/10.1016/j.chb.2022.107308

Corvelo Benz, N. L. & Gomez Rodriguez, M. (2024). Human-Aligned Calibration for AI-Assisted Decision Making. Preprint, *arXiv*. https://doi.org/10.48550/arXiv.2306.00074

Crootof, R., Kaminski, M. E., & Price, W. N. II. (2023). Humans in the loop. *Vanderbilt Law Review, 76*(2), 429–510. https://heinonline.org/HOL/PDF?handle=hein.journals/vanlr76&id=447

Defense Advanced Research Projects Agency. (n.d.). In the Moment. https://www.darpa.mil/research/programs/in-the-moment. Accessed June 25, 2025.

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing, *Journal of the Academy of Marketing Science, 48*(1), 24-42. https://doi.org/10.1007/s11747-019-00696-0

Davis, F. D. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIT Quarterly, 13*(3), 319-339. https://doi.org/10.2307/249008

Dingel, J., Kleine, A., Cecil, J., Sigl, A., Lermer, E., & Gaube, S. (2024). Predictors of Healthcare Practitioners' Intention to Use AI-Enabled Clinical Decision Support Systems (AICDSSs): A Meta-Analysis Based on the Unified Theory of Acceptance and Use of Technology (UTAUT). Preprint, *Journal of Medical Internet Research.* https://preprints.jmir.org/preprint/57224

Dodig Crnkovic, G., Basti. G. & Holstein, T. (2025). Delegating Responsibilities to Intelligent Autonomous Systems: Challenges and Benefits. *Bioethical Inquiry.* https://doi.org/10.1007/s11673-025-10428-5

Dominguez-Martinez, S., Sloof, R., & Von Siemens, F. A. (2014). Monitored by your friends, not your foes: Strategic ignorance and the delegation of real authority. *Games and Economic Behavior, 85*(1) (2014), 289-305. https://doi.org/10.1016/j.geb.2014.02.003

Fuchs, A., Passarella, A., Conti, M. (2025). Optimizing Delegation Between Human and AI Collaborative Agents. In: Meo, R., Silvestri, F. (eds) Machine Learning and Principles and Practice of Knowledge Discovery in

Databases. ECML PKDD 2023. Communications in Computer and Information Science, vol 2134. Springer, Cham. https://doi-org.umasslowell.idm.oclc.org/10.1007/978-3-031-74627-7_18

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Cognitive challenges in human-AI collaboration: Investigating the path towards productive delegation. Information Systems Research, *33*(2), 678-696. https://doi.org/10.1287/isre.2021.1079

Gevorgyan, M. (2023). Crossing Boundaries: Autonomous Weapon Systems and the Challenge of IHL Compliance. *Int'l JL Ethics Tech.*, 21.

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the Technology Acceptance Mode; to assess automation. *Cognition Technology and Work, 14*(39), 49. https://doi.org/10.1007/s10111-011-0194-3

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627-660. https://doi.org/10.5465/annals.2018.0057

Hales, C. (1999). Leading horses to water? The impact of decentralization on managerial behaviour. *Journal of Management Studies, 36,* 831–851. https://doi.org/10.1111/1467-6486.00160

Hemmer, P., Westphal, Monika, Schemmer, M., Vetter, S., Vössing, M., & Satzger, G. (2023). Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia*. ACM, New York, NY, USA, 11 pages. https://www.researchgate.net/deref/https%3A%2F%2Fdoi.org%2F10.1145%2F3581641.3584052?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Fenster, M., Zuckerman, I., & Kraus, S. (2012). Guiding user choice during discussion by silence, examples and justifications. *Frontiers in Artificial Intelligence and Applications*, *242*, 330–335. https://doi.org/10.3233/978-1- 61499-098-7-330

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human–AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Johnson, J. (2020). Delegating strategic decision-making to machines: Dr. Strangelove Redux? *Journal of Strategic Studies, 45*(3), 439–477. https://doi.org/10.1080/01402390.2020.1759038

Koenig, P. D. (2025). Attitudes toward artificial intelligence: combining three theoretical perspectives on technology acceptance. *AI & Society, 40*, 1333–1345. https://doi.org/10.1007/s00146-024-01987-z

Leana, C. R. (1986). Predictors and consequences of delegation. *Academy of Management Journal, 29*(4), 754–774. https://doi.org/10.2307/255943

Lee, K. M., Peng, W., Jin, S. A., & Yan, C. (2006). Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human-robot interaction. *Journal of Communication, 56*(4), 754–772. https://doi.org/10.1111/j.1460-2466.2006.00318.x

Leetaru, K. (2016, June 18). *Will AI and robots make humans obsolete?* Forbes. https://www.forbes.com/sites/kalevleetaru/2016/06/18/will-ai-and-robots-make-humans-obsolete/

Lupia, A. (2015). Delegation of power: Agency theory, In James D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (pp. 58-60), Elsevier.

MacCrory, F., Westerman, G., Alhammadi, Y., & Brynjolfsson, E. (2014). *Racing with and against the machine: Changes in occupational skill composition in an era of rapid technological advance* [Working paper]. https://www.researchgate.net/publication/264696560

McVay, J. (2025, June). DARPA In the Moment. Presented at the 2025 Human Alignment in AI Decision-Making Systems: An Inter-disciplinary Approach towards Trustworthy AI, IEEE CAI 2025 Workshop, Santa Clara, CA.

National Security Commission on Artificial Intelligence (NSCAI). 2021. *National Security Commission on Artificial Intelligence: Final Report*. 1 March. Washington, DC: NSCAI. Available from: www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.

National Science and Technology Council. (2023). *National artificial intelligence research and development strategic plan: 2023 update (NITRD Publication).* Select Committee on Artificial Intelligence, Executive office of the President of the United States. https://www.nitrd.gov/pubs/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf

Owens, D., Grossman, Z., & Fackler, R. (2014). The control premium: A preference for payoff autonomy. *American Economic Journal: Microeconomics, 6*(4), 138–161. doi:10.1257/mic.6.4.138.

Pedro, A. R., Dias, M. B., Laranjo, L., Cunha, A. S., & Cordeiro, J. V. (2023). Artificial intelligence in medicine: A comprehensive survey of medical doctor's perspectives in Portugal. *PloS One, 18*(9), e0290613. https://doi.org/10.1371/journal.pone.0290613

Power, N., & Alison, L. (2017). Redundant deliberation about negative consequences: Decision inertia in emergency responders. *Psychology, Public Policy, and Law, 23*(2), 243–258. https://doi.org/10.1037/law0000114

Richardson, H. A., A. C. Amason, A. K. Buchholtz and J. G. Gerard. (2002). CEO willingness to delegate to the top management team: The influence of organizational performance. *International Journal of Organizational Analysis, 10*(2), 134–155. https://psycnet.apa.org/doi/10.1108/eb028947

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). *Would you trust a (faulty) robot?* Proceedings of the Tenth Annual ACM/IEEE International Conference on Human- Robot Interaction - HRI '15, 141–148. https://doi.org/10.1145/2696454.2696497

Shortland, N. D., & Alison, L. J. (2020). Colliding sacred values: a psychological theory of least-worst option selection. *Thinking and Reasoning, 36*(1), 1818 – 139. https://doi.org/10.1080/13546783.2019.1589572

Shortland, N.D., Alison, L. J., & Moran, J. (2019). *Conflict: How Soldiers Make Impossible Decisions*. New York: Oxford University Press.

Shortland, N. D., Alison, L. J., & Thompson, L. (2020). Military Maximizers: Examining the effect of Individual Differences in Maximization on Military Decision Making. *Personality and Individual Differences, 163*. 110051. https://doi.org/10.1016/j.paid.2020.110051

Shortland, N.D., Alison, L., & Barrett-Pink, C. (2018). Military (in)decision-making process: a psychological framework to examine decision inertia in military operations. *Theoretical Issues in Ergonomics Science, 19*(6), 752–772. https://doi-org.umasslowell.idm.oclc.org/10.1080/1463922X.2018.1497726

Song, J. & Lin, H. (2023). Exploring the effect of artificial intelligence intellect on consumer decision delegation: The role of trust, task objectivity, and anthropomorphism. *Journal of Consumer Behavior, 23,* 727-747. https://doi.org/10.1002/cb.2234

Spitzer, P., Holstein, J., Hemmer, P., Vössing, M., Kühl, N., Martin, D. & Satzger, G. (2025). Human Delegation Behavior in Human-AI Collaboration: The Effect of Contextual Information. In *Proceedings of the ACM on Human-Computer Interaction, 9*(2), 1-28. https://doi.org/10.1145/3710999

Steffel, M., E. F. Williams and J. Perrmann-Graham. (2016). Passing the buck: Delegating choices to others to avoid responsibility and blame. *Organizational Behavior and Human Decision Processes, 135*, 32–44. https://doi.org/10.1016/j.obhdp.2016.04.006

Stevens, A. F., & Stetson, P. (2023). Theory of Trust and Acceptance of Artificial Intelligence Technology (TrAAIT): An Instrument to Assess Clinician Trust and Acceptance of Artificial Intelligence. *Journal of Biomedical Informatics*, *148*, 104550. https://doi.org/10.1016/j.jbi.2023.104550

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science, 185*(4157), 1124-1131.

van den Heuvel, C., Alison, L., & Crego, J. (2012). How uncertainty and accountability can derail strategic 'save life' decisions in counter-terrorism simulations: a descriptive model of choice deferral and omission bias. *Journal of Behavioral Decision Making, 25*(2), 165–187. https://doi.org/10.1002/bdm.723

Waring, S., Alison, L., Cunningham, S. & Whitfield, K. C. (2013). The impact of accountability on motivational goals and the quality of advice provided in crisis negotiations. *Psychology, Public Policy and Law, 19*(2), 137-150. https://doi.org/10.1037/a0030629

Weinberger, D. (2019). *Everyday chaos: Technology, complexity, and how we're thriving in a new world of possibility.* Harvard Business Review Press. https://books.google.com/books?id=R7V2DwAAQBAJ

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83*(2), 260-281. https://doi.org/10.1006/obhd.2000.2909

Yukl, G. & P. P. Fu. (1999). Determinants of delegation and consultation by managers. *Journal of Organizational Behavior*, *20*(2), 219–232.

Zhang, Y., Liao, Q., & Bellamy, R. (2020). Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3351095.3372852