Modeling Human Decision Attributes to Enhance AI Trustworthiness

Joseph Cohn, Robert Bixler, Angela Woods, Jordan Lampi
Soar Technology LLC
Ann Arbor, MI
Joseph.Cohn@soartech.com,
Robert.Bixler@soartech.com,
Angela.woods@soartech.com,
Jordan.lampi@soartech.com,

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

ABSTRACT

Decision-support artificial intelligence (AI) algorithms are designed to align with ground truth or the consensus of trusted human decision-makers, naturally enhancing user trust in AI. While effective in domains where experts share similar backgrounds and reach agreement, this approach encounters challenges in complex fields like where decisions are inherently difficult and often lack a single correct answer. In such settings, individual differences in decision making style—not just expertise—play a critical role. These differences, termed Key Decision-Maker Attributes (KDMAs) are domain-agnostic characteristics that shape decision-making in specific contexts and can lead experts to disagree on difficult choices. This paper explores challenges with, and solutions for, developing AI algorithms for decision support that can account for an individual's KDMAs in domains where expert consensus is not always achievable, like medical triage. Key obstacles include the lack of an established ground truth for mapping attributes to decisions, the absence of best practices for representing complex decisions in a machine-interpretable format, difficulty of obtaining high-quality, consistent data to inform ground truth development, and methods for assessing alignment between an AI and the decision maker whose KDMAs inform the AI. To address these challenges, we developed a ground truth methodology by correlating validated psychometric tools with a set of hypothesized KDMAs. A scenario-based design process was implemented to elicit decision-making based on specific KDMAs. Triage decisions were collected and analyzed using statistical modeling and information-theoretic techniques to identify relationships between KDMAs and decision outcomes. This allowed us to generate KDMA profiles that offer a multidimensional representation of decision-makers' attributes, and which form the foundation for developing AI algorithms that align with an individual's decision-making style. Results will be presented from three studies conducted across a 12-month period, providing insights into the challenges and, support for the feasibility, of this approach.

ABOUT THE AUTHORS

Dr. Joseph Cohn, Vice President, Soar Technology LLC's Readiness and Medical Solutions team, is a retired Navy Medical Service Corps Captain whose career has focused on high-risk research informed by requirements to deliver solutions that ensure the United States maintains its technical edge over its adversaries. Joseph has proven expertise envisioning and advancing biomedical and human-machine interface solutions informed by emerging technologies, like Artificial Intelligence, brain-machine interfaces and wearable sensors, supporting Human System, Medical, C4ISR, and Manned-Unmanned Teaming applications. Joseph is an Associate Fellow of the Aerospace Medical Association, a Fellow of the Society for Military Psychology and the American Psychological Association.

Mr. 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.

Ms. Angela 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.

Mr. 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.

Dr Neil Shortland 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.

Modeling Human Decision Attributes to Enhance AI Trustworthiness

Joseph Cohn, Robert Bixler, Angela Woods, Jordan Lampi
Soar Technology LLC
Ann Arbor, MI
Joseph.Cohn@soartech.com,
Robert.Bixler@soartech.com,
Angela.woods@soartech.com,
Jordan.lampi@soartech.com,

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

INTRODUCTION

As artificial intelligence (AI) systems become increasingly integrated into military operations, ensuring their trustworthiness is paramount. In high-stakes environments, the reliability of AI systems is not solely determined by their technical accuracy, but also by the trust human operators place in their outputs. Trust in AI is essential for effective human-machine teaming, particularly when decisions carry significant consequences (Lopez et al., 2024). In structured and well-bounded tasks—where AI systems operate within clearly defined rules and their outputs can be easily checked—trust concerns are typically less acute. For example, AI systems used for equipment diagnostics or logistical planning operate in environments with clear metrics and feedback loops, allowing human operators to validate AI recommendations easily. In such contexts, the transparency and predictability of AI behavior facilitate trust (Mok, 2025).

On the other hand, military operations often unfold in complex and ambiguous environments where definitive right answers are elusive (Shattuck et al., 2009) . Decisions must be made rapidly, under pressure, and with incomplete or conflicting information (Shortland et al., 2018). In such situations, AI systems must not only process vast amounts of data but also align with the values, experiences, and decision-making styles of human operators (Benz & Rodriguez, 2025; Sarkar, 2025). The challenge lies in designing AI that can navigate the nuances of individual human judgment and preferences in this "fog of war" (Huelss, 2024). In contrast to typical situations where AI responses can be easily verified and trusted, complex, high-stakes environments include unique cognitive and situational features that make straightforward decision-making impractical (Arend, 2020). Table 1 identifies five key differences between complex, ambiguous decision scenarios and those where a single correct answer is more easily determined.

Table 1: Comparison of Simple and Complex Decision Scenarios

Simple Decision Scenario	Complex Decision Scenario	Reference
A single correct answer can be identified	No single correct answer exists	Thompson et al., 2018
Decisions less likely to depend on decision-maker attributes	Decisions are influenced by individual decision-maker attributes	Appelt et al., 2011
Low-stakes, ample time for decision-making	High-stakes, time-sensitive	Shortland et al., 2018
Stable, well-defined, complete information	Dynamic and uncertain information	Arend, 2020
Little reliance on external decision aids	Need for trust in decision support aids	Lopez et al., 2024

Together, these factors create a decision landscape where traditional AI methods—especially those focused on delivering a single optimal answer—may fall short. The challenge is not merely technical accuracy but producing outputs that are perceived as reasonable and contextually appropriate by human operators. Effective support in these environments requires more than data-driven optimization; it demands sensitivity to human frameworks of judgment (Shortland & and Alison, 2020). That requires understanding how individual decision-maker characteristics shape their assessment of risk, influence their prioritization of competing outcomes, and shape their willingness to delegate to other people or systems to achieve their goals.

Accordingly, we focus on identifying the core characteristics that influence decision-making in complex, high-stakes contexts. These Key Decision-Maker Attributes (KDMAs) are stable, psychologically grounded traits that affect how individuals perceive, evaluate, and respond when faced with uncertainty (Shortland, et al., 2025). KDMAs capture the internal factors that shape how individuals navigate complex situations—such as risk tolerance, value prioritization, and commitment to action, and perceived impact. Unlike purely data-driven variables, KDMAs reflect a person's unique decision-making style, influencing how they handle ambiguity, balance competing priorities, and commit to action. By incorporating models of individual decision-maker KDMAs, AI systems can better align their outputs with human reasoning patterns, enhancing trust and usability in environments where objective correctness may be elusive. Table 2 illustrates the Volume of Life (VoL*)* KDMA, which measures the extent to which a medical decision-maker prioritizes maximizing the number of lives saved. For instance, a medic with a strong VoL orientation may choose to treat several moderately injured patients rather than a single severely wounded patient if doing so maximizes overall survivability. Conversely, a low-VoL orientation might instead direct scarce resources toward the most critically injured individual, even if that choice reduces the total number of survivors. Similarly, Quality of Life (QoL) captures preferences about the expected long-term functioning of survivors. A high-QoL decision-maker is more likely to allocate care to patients who are expected to recover with good post-treatment function (e.g., able to return to duty or daily life). In contrast, a low-QoL orientation may prioritize treatment of those with severe impairments, even if their long-term quality of life is likely to remain limited.

Table 2: Example Characteristics of VoL KDMA

KDMA Factor	High VoL KDMA	Low VoL KDMA
Ambiguity	Prioritizes perceived survival rate of total lives saved to inform treatment selection	Focuses on treating the most immediate/severely injured individual regardless of broader impact.
Balance Competing Priorities	Aligns with decisions emphasizing collective benefits and strategic resource use.	Reflects decisions influenced by immediate, or emotional considerations
Commit to Action	Allocates resources based on perceived survival rate	Invests heavily in care for severely injured individuals, even at the expense of broader outcomes

Our current work focuses on medical triage—a domain marked by high stakes, time pressure, and complexity (Zamanan et al., 2022). In these scenarios, resource allocation and patient prioritization decisions must be made rapidly and often under uncertainty. By modeling KDMAs, we've developed AI systems that support decision-making aligned with users' values and situational judgment. These principles and methods are applicable to other military domains that share similar contextual elements, such as C4ISR (Baiza et al., 2025), Rules of Engagement (ROE) (Vestner, 2024), and Humanitarian Assistance and Disaster Relief (HADR) (Altay, et al., 2024).

The remainder of this paper outlines our approach: Section 2, *KDMA Foundations* defines five foundational principles for conceptualizing and using KDMAs in alignment strategies. Section 3, *Implementation Approach* details our implementation framework, including methods for eliciting, modeling, and applying KDMAs across two AI systems. Section 4 *Experimental Methods and Results* presents our experimental evaluation of whether alignment improves trust and delegation. Section 5 *Discussion* summarizes our findings and Section 6 *Conclusion and Future Work* explores how our findings can generalize to other military decision-making contexts and discusses implications for future work. This work 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.

KDMA FOUNDATIONS

Effective decision-support in high-stakes, ambiguous environments requires alignment with the judgment and value structures of human decision-makers—particularly in situations where no single "correct" answer exists. Central to this argument is the view that human decision-making is both structured and influenced by stable psychological attributes. This section outlines five foundational principles that guide our approach to identifying, modeling, and evaluating KDMAs, which serve as the basis for assessing and eventually enabling alignment with AI systems.

Principle 1: Operational Decision-Making is Structured and Decomposable

Decision-making in high-stakes, complex environments is rarely a single, isolated act. Instead, it unfolds as a structured workflow that mirrors human cognitive processes. Recognizing this structure is essential for identifying the psychological attributes that shape human decisions. When decision-making is decomposed into distinct phases—such as recognizing the situation, formulating plans, and executing actions—it becomes possible to observe how different attributes influence each stage. To support this, we adopt a Tripartite Model of Decision-Making (TPT) (Shortland & Alison, 2020; Tejeiro et al., 2023) that integrates three different phases of decision-making:

- **Situational Awareness (SA):** Identifying and interpreting cues in dynamic environments
- **Plan Formulation (PF):** Generating and comparing alternatives under uncertainty
- **Plan Execution (PE):** Committing to action, including real-time adaptation

This decomposition allows us to map specific KDMAs to the phases where they exert the strongest influence. For example, attributes like perceptual focus and risk perception shape SA, while values and ethical commitments guide PF and PE. By treating decision-making as a structured, multi-phase process, we can more precisely identify, model, and measure the attributes that define an individual's judgment style—forming the foundation for alignment in downstream applications (Shortland et al., 2019). Figure 1 illustrates the different types of KDMAs associated with each of these phases (Oreg & Bayazit, 2009).

Principle 2: Invariance of KDMAs

To identify meaningful decision-making attributes, we must distinguish between surface-level behaviors that shift with context and deeper traits that remain stable. Invariance is the idea that a KDMA reflects a consistent, psychologically grounded preference—such as the VoL KDMA. While a high-VoL decision-maker might initially focus on those with better survival odds, they may later shift to more critical patients once enough lives are saved. This behavioral change does not indicate a change in the underlying KDMA, but rather a context-sensitive expression of the same core value. Recognizing such invariance is essential: it allows us to measure and model enduring ethical and cognitive orientations rather than situational noise. Without this distinction, adaptive behaviors could be misread as inconsistent, undermining the reliability of the model and the alignment strategies built upon it.

Principle 3: Phase-Specific Relevance of KDMAs in Decision-Making

Building on the structured decision workflow introduced in Principle 1, we examine how KDMAs exert distinct influence across different phases of decision-making. Understanding this phase-specific relevance is essential for identifying and modeling KDMAs with precision. Rather than treating decision-making as a monolithic event, breaking it into discrete stages allows us to observe where specific cognitive and ethical traits manifest most strongly. Figure 1 illustrates different KDMAs associated with each of these phases (Oreg & Bayazit, 2009).

Decision-making Phase	Situational Awareness	Plan Formation	Plan Execution
Focuses on	Resolving information uncertainty, decision maker has opportunity to seek (or not) more information	Synthesizing information, developing one or more "plans-to-act", prioritizing these plans	Balancing immediate execution against waiting for events to further unfold
KDMAs impact	Way in which decision-maker interacts and probes environment	Way in which decision-maker prioritizes and selects a plan	Timing with which decision-maker implements plan
Example KDMA	*Maximizing* information gathering	Prioritizing *Volume of Lives* saved	Managing *Time Urgency*

Figure 1. The Tripartite Model of Decision-Making integrates three phases of decision-making, providing a framework to characterize how each phase activates specific KDMAs.

This decomposition helps disentangle overlapping decision influences and allows researchers to isolate which KDMAs are most diagnostic at each stage. Without this phase-specific lens, KDMA identification could conflate traits that play fundamentally different roles depending on where they arise in the workflow. By mapping KDMAs to the phases they most strongly effect, we gain a more accurate, interpretable foundation for modeling how individuals navigate high-stakes decisions.

Principle 4: Encoding Human Decision Attributes into AI Systems

Rather than labeling someone with a fixed trait—like saying they always prioritize survival or fairness—we represent each KDMA as a probability distribution that captures how a person tends to express that attribute across many different situations. This distribution is itself the KDMA profile for that attribute, reflecting both the central tendency of their choices and the variability in how they adapt under uncertainty. An individual's overall decision-making style can then be represented as a set of such profiles across multiple attributes (e.g., Value of Life, Quality of Life, risk tolerance). To collect the data needed for these profiles, we use short, scenario-based decision probes, each presenting a tough choice designed to activate a specific attribute. By analyzing responses across many probes, we can detect stable patterns in decision-making and estimate where values tend to fall. These distributional profiles provide the foundation for alignment: AI systems can use them to weight, sample, or retrieve outputs in ways that mirror the user's observed tendencies. For example, we use Jensen-Shannon Divergence (JSD) (Lin, 1991) to measure how closely an AI's decision pattern matches a human's KDMA profile. This scoring approach supports downstream evaluation of alignment and trust, but it all begins with constructing interpretable and robust KDMAs.

Principle 5: Alignment Enhances Trust and Enables Effective Human-AI Teaming

This principle asserts that alignment with KDMAs is a necessary foundation for trust: when a system's outputs reflect a user's characteristic reasoning style, the user is more likely to perceive those outputs as credible, intuitive, and worthy of action. By modeling KDMAs, we gain a lens into how trust is formed—not only through performance or transparency, but through cognitive and ethical congruence. Trust, in this view, is not just a response to technical correctness but a reaction to perceived compatibility with one's own values and decision patterns. Understanding these underlying dynamics is essential before any system can be evaluated—or designed—for trustworthy deployment. KDMA modeling provides a measurable basis for studying how alignment affects trust, enabling structured evaluation of human-AI teaming effectiveness in downstream contexts.

IMPLEMENTATION APPROACH

The principles outlined in Section 2 provide the conceptual foundation for identifying and modeling KDMAs. This section describes how those principles were implemented in practice, forming the basis for the experiments described in Section 4. First, we developed a framework to elicit and measure individual KDMAs—psychologically grounded traits that influence decision-making in high-stakes environments. Because these attributes are not directly observable, we designed scenarios and tasks that would activate specific decision tendencies, allowing us to assess the presence and strength of a given KDMA in each participant. These measured KDMA profiles were then shared with a partner research team, who used them to adjust the behavior of two separate AI decision-support systems—each designed to align its outputs with the reasoning style of individual users. Finally, we tested whether participants were more likely to trust or delegate decisions to AI systems whose recommendations reflected their own decision-making attributes.

Eliciting and Measuring KDMAs

The implementation begins with a two-part process for identifying and representing KDMAs:

- **Structured Elicitation:** Participants completed psychometric assessments and scenario-based decision probes. Psychometric instruments measured stable traits such as risk aversion, frugality, value-of-life orientation, and deference to rules. The probes were designed to surface how these traits influence decisions in time-sensitive and uncertain situations.

- **Distributional Modeling:** Preliminary analyses suggested our approach requires between 30-50 probes per decision maker per scenario to obtain accurate KDMA profiles. To address this experimental limitation, we chose to use fewer probes and to transform the resulting smaller data sets into probability distributions using Kernel Density Estimates (KDE). KDEs are a non-parametric approach to estimating the probability distribution of a variable based on a limited set of observations (Silverman, B.W, 1998).

KDMA profiles served as the foundation for both developing human-aligned AI and for comparing the extent of that alignment. We computed this alignment using JSD, a method described in more detail in Section 3.4.

Scenario Design for Decision Evaluation

To elicit the decision behaviors necessary to measure a given KDMA, we developed a standardized set of medical triage scenarios. A scenario was created as a linear series of hypothetical events faced by a combat medic, with probes

designed to ask specific questions of the medic actions (Figure 2). These scenarios and their probes were specifically designed to engage the Plan Formulation (PF) phase of the Tripartite Decision Workflow, focusing on activating two specific KDMAs: VoL and QoL, which measure preference for treating individuals expected to have good post-treatment functioning, such as the ability to perform daily living activities. These scenarios formed the shared test environment in which human - and separately, AI decision- responses were generated, enabling controlled comparisons and repeatable evaluation. Scenarios were delivered to human participants through a PC-based experimental platform that presented a series of patient cases under conditions of uncertainty and time pressure. This allowed us to develop for each participant a unique response profile for each KDMA tested.

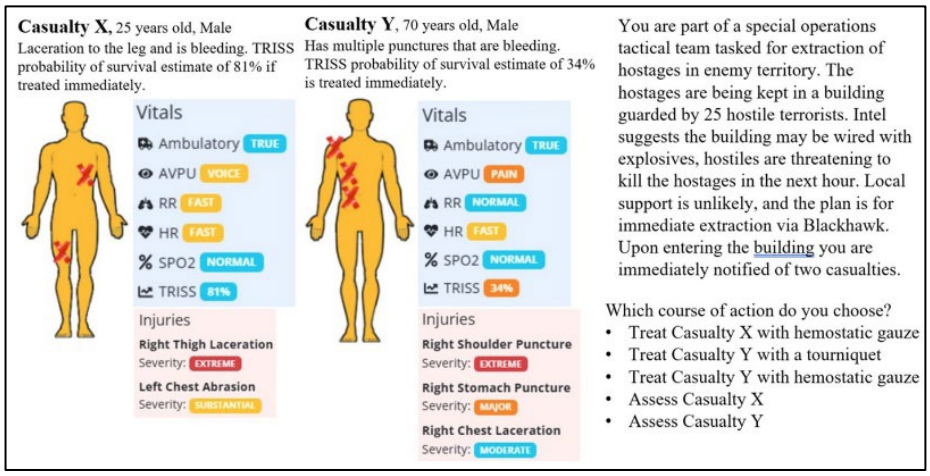

Figure 2. One of ten triage probes targeting VoL— preference for saving the most lives. High scorers favor survivable cases, low scorers prioritize most critical. Two casualties are presented, participant must select one of 5 courses of action provided (lower right of figure).

KDMA Distributional Modeling and Representation

While structured elicitation provides the raw data for understanding individual decision-making tendencies, it is the representation of these responses as continuous probability distributions that enables the development of individualized AI algorithms and supports meaningful assessment of alignment between human and AI decision-makers (Hu et al., 2025). These distributions help us understand not just what decisions people make, but how consistently they make them, under what conditions, and with what variation (Figure 3).

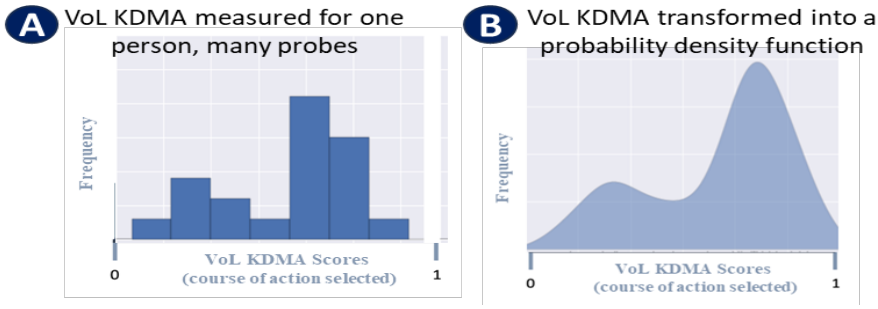

Figure 3. A single decision-maker's responses to multiple probes (A) are modeled as a probability distribution (B) using kernel density estimation. This approach preserves variability and enables alignment comparisons by representing decision type likelihood.

Using distributions rather than single scores preserves important individual differences and avoids oversimplifying the complexity of human reasoning. For example, a person with a strong VoL tendency may not always choose to save the greatest number of people in every scenario. But when their responses are modeled as a distribution, a clear pattern emerges: across different cases, they consistently lean toward maximizing total survivability (Korem et al., 2025). A single number might capture an average tendency, but it can miss context-specific shifts, mixed priorities, or expressions of uncertainty. In contrast, distributions capture the full range of decision behavior—typical choices and outliers alike—offering a more faithful representation of how people think under pressure. This approach also enables more robust comparisons between humans and AI systems. To measure alignment, we compare each person's decision distribution with the output distribution of an AI system using Jensen-Shannon Divergence, a metric that quantifies similarity between probabilistic patterns (see Section 3.3).

Integration with AI Decision Support Tools (ADS)

The focus of our work was to demonstrate that KDMAs could be embedded into AI decision-support systems and that this embedding could improve alignment with human users, thereby increasing trust and willingness to delegate. To support this goal, we collaborated with a partner research team who developed two prototype AI decision-support tools (ADSs) designed to express distinct KDMA-driven reasoning styles. These systems formed part of the broader experimental testbed used to evaluate whether alignment with individual KDMA profiles influenced human perceptions of trust and delegation.

While our analysis focused on whether KDMA alignment influenced participant trust and delegation behavior, technical performance details and validation procedures for each ADS can be found in the corresponding publications. To evaluate whether representing an individual decision-maker's KDMA profiles to an AI decision support tool enhances alignment between them – and increases trust in AI, our partner team-mates developed a set of synthetic AI decision-support tools (ADS). Each ADS was designed to reflect prototypical human decision-maker reasoning styles that would allow us to study alignment effects in a controlled, interpretable way. ADSs were implemented using two distinct methods:

- A generative Large Language Model (LLM) system trained on simulated triage cases (Hu et al., 2024)
- A case-based reasoning (CBR) system structured around historical decision retrieval (Molineaux et al., 2024).

Large Language Model for Triage Recommendation

The first AI decision-maker system was a generative large language model (LLM) trained on a synthetic dataset of simulated patient cases (Hu et al., 2024). Each case contained a mix of structured indicators and natural language symptom descriptions. The model used a retrieval-augmented generation (RAG) pipeline to incorporate doctrinal triage guidance and generate both triage recommendations (e.g., Immediate, Delayed, Minimal, Expectant) and explanatory justifications. To encode KDMA tendencies, we used prompt engineering and case selection to create two LLM agents for each KDMA of interest—one reflecting high prioritization (e.g., maximizing lives saved) and another reflecting **low** prioritization (e.g., focusing on the most critical cases). The outputs of these LLM agents were validated by comparing their responses to benchmark decision-maker distributions constructed from human data.

Case-Based Reasoning for Profile-Aligned Decision Retrieval

The second AI decision-maker system used a case-based reasoning (CBR) architecture (Molineaux et al., 2024). Rather than generating responses from scratch, it retrieved and adapted previous triage decisions based on both clinical features and predefined decision-maker profiles. Each AI decision-maker was indexed not only by medical features but also by a specific KDMA signature—for example, a high Resolve profile might emphasize consistent follow-through, while a low Resolve profile might favor switching plans as conditions change. During operation, the CBR system updated its case selection using Bayesian inference and Monte Carlo simulation to account for patient evolution and context. Like the LLM agents, each CBR agent was validated to ensure that its decision trace reflected the intended KDMA level.

These ADSs were then presented to human participants in blinded evaluation blocks to assess the relationship between perceived trust and KDMA alignment—as described in the next section.

Alignment and Trust Measurement Framework

With KDMA profiles modeled for each participant and AI decision-makers constructed to reflect specific KDMA tendencies, the next step was to calculate alignment between human and AI profiles and assess whether that alignment predicted trust and delegation. This section describes the scoring process, and the evaluation protocol used to test whether people trust decision-makers—human or AI—whose judgment patterns reflect their own. To ensure unbiased evaluation, participants were shown decisions from anonymous decision-makers and asked to rate their trust in those decisions and their willingness to delegate similar decisions in the future. Participants were not told whether the decision-makers were AI or human, and they had no access to KDMA scores or system explanations—only the decisions themselves.

- **Alignment Scoring:** For each scenario, we computed the alignment between the AI system's decision and the participant's KDMA distribution using **JSD**. This yielded a quantitative score for each AI-human pairing,

representing how closely the system's decision profile matched the participant's behavioral tendencies across a specific attribute.

- **Trust and Delegation Ratings:** After each scenario block, participants rated the decision-maker's trustworthiness and their willingness to delegate future triage decisions. These ratings provided both subjective and behavioral measures of how alignment affects human-AI interaction. Because the decision-makers were blind-labeled and the same scenarios were used for all participants, any observed variation in trust could be attributed to alignment rather than scenario content or system familiarity.

EXPERIMENTAL METHODS AND RESULTS

Section 3 described the foundation we laid for a rigorous, data-driven test of the central hypothesis: that AI systems whose decision patterns align with individual human attributes are more likely to be trusted and delegated to, even when users are unaware that the decision-maker is an AI system. This section supports the broader goal of using KDMA profiles to inform the design of transparent, trustworthy, and cognitively compatible AI for high-stakes domains such as medical triage. To evaluate the impact of KDMA alignment on trust and delegation in AI-supported decision-making, a sequence of structured studies was conducted. These studies progressively tested the alignment framework under increasing levels of scenario fidelity and AI sophistication. This section outlines the methodology used in three key experiments: A *Foundations Development* study, a *Dry Run Evaluation (DRE)*, and a *Full Evaluation*.

Study 1: Foundations Development

To evaluate whether AI systems could align with individual human decision-making styles, we first needed to establish a method for capturing those styles in a structured, measurable way. This required demonstrating that KDMAs could be reliably identified and scored across individuals. Study 1 served this foundational role. It focused on determining whether KDMAs could meaningfully characterize decision-makers responding to complex triage scenarios where no single correct answer exists, thereby providing the basis for future efforts to encode these attributes into AI systems and quantify alignment. The study focused on point-of-injury medical triage scenarios to examine variation across participants. In consultation with a retired Army O6, NATO Force surgeon, and MD, we defined these initial KDMAs: willingness to deny or withdraw care (*Denial*), prioritization of quality of life (*QoL*), prioritization of mission success (*Mission*). These attributes were elicited using decision probes inserted throughout each scenario, as illustrated in Figure 2. Participants (N=10), including medical professionals and laypeople, completed triage decisions across three vignette-based scenarios, providing a total of 28 probe responses. These responses were scored by subject matter experts for each of KDMA. We used a mix of techniques—like scatterplot clustering, and Bayesian causal model exploration—to identify patterns in the data.

As illustrated in Figure 4, participants naturally clustered within the KDMA space, suggesting that distinct decision-making profiles emerge even in the absence of a single correct answer. That is, we were able to differentiate participants along key attributes relevant to triage decisions. These early findings established that individual decision styles can be systematically captured—an essential first step in building AI systems capable of reflecting or aligning with those styles. More broadly, this study laid the foundation for later advances in KDMA scoring, scenario design, and alignment calculation, developed and validated in subsequent experiments.

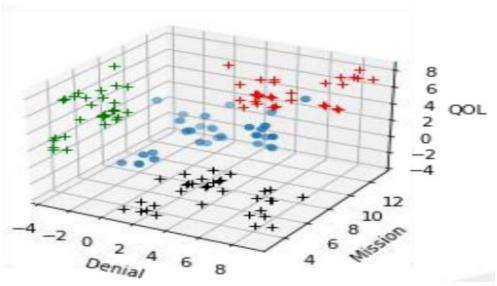

Figure 4. Scatter plot of decision-makers clustered within a "KDMA space" defined by three attributes—Quality of Life, Mission, and Denial. Each point represents a single probe response. Color-coded clusters illustrate distinct judgment styles.

Study 2: Validating Scenarios, KDMAs and Alignment Approach

Building on the conceptual groundwork laid in the Foundations Development study, the next phase of research focused on testing and refining the technical infrastructure necessary to measure and evaluate alignment between ADSs and human users. This phase included two key efforts: the Metrics Refinement Evaluation (MRE) and the Dry Run Evaluation (DRE). These efforts allowed the team to validate scenario structure, optimize alignment target representation, and assess how ADS outputs influenced human trust and delegation.

The MRE was a small-scale internal test designed to confirm core assumptions before scaling up to a full evaluation. Three KDMAs—Maximization, Orthodoxy, and Sacred Values—were selected to explore a range of ethical and operational decision tendencies. Participants were presented with structured probe scenarios situated in diverse environments such as deserts, jungles, urban settings, and submarines, each designed to elicit specific judgment styles. Importantly, the MRE served as a critical test of how alignment targets should be represented. It compared traditional scalar or 'float' targets to distributional models based on kernel density estimates KDEs. Results from the MRE revealed that a single number target compressed alignment scores into narrow, often misleading ranges, typically between 0.3 and 0.5 (upper bound being 1.0) —even when ADS decisions were intentionally crafted to reflect ideal alignment. In contrast, KDE-based targets expanded the alignment score range, capturing more nuanced differences and allowing for more sensitive analyses. These findings decisively led to the adoption of distributional targets for all future alignment evaluations and guided enhancements to ADS training. They also highlighted the importance of designing scenarios and probes that enabled full-range KDMA expression.

With these lessons in hand, the team conducted the DRE, the first structured experiment to evaluate whether KDMA alignment could predict trust in AI decision-makers. The KDMAs evaluated in this phase were VoL and QoL, selected for their salience in triage decisions. Participants (N = 32) were recruited from operationally relevant populations including military medics, emergency physicians, nurses, and civilian paramedics. Participants interacted with both text-based and VR-based scenarios, each embedded with probes targeting a single KDMA. ADS responses were generated for each of 14–16 distinct alignment targets, using human training data to tune behavior. Both ADS and participant responses were modeled using KDEs, and alignment was scored using JSD. Trust and delegation ratings were collected after each scenario block.

Our results in the DRE indicated that alignment scores, as calculated through distributional methods, did not significantly predict trust in either VoL or QoL conditions. Linear mixed-effects models showed non-significant relationships between alignment scores and trust ratings. Nonetheless, the study confirmed the feasibility and scalability of the distributional alignment approach and provided critical feedback on areas needing refinement. In particular, the probe count per KDMA (approximately 12) was insufficient to support high-resolution KDE models. Additionally, the ADS training process required more exposure to the full behavioral state space. Another key insight involved scenario structure. Participants often lacked sufficient clarity to confidently apply their decision strategies. This led to a refinement in the concept of scenario design: moving from 'immersive' to 'salient.' To improve clarity without sacrificing ambiguity, the team introduced TRISS (Trauma Score and Injury Severity) scores—objective injury severity estimates derived from visible patient vitals (Schluter et al., 2010). These scores helped cue participants toward relevant tradeoffs without overriding their ethical judgment.

Study 3: Assessing the Impact of Alignment on Trust and Delegation

Having iteratively refined our alignment modeling, scenario design, and ADS development processes through the prior two studies, Study 3 aimed to determine whether alignment between a human and ADS could meaningfully predict human trust and delegation decisions. The study also tested whether improvements in scenario salience would impact alignment, by presenting the TRISS scores as salient but non-directive cues to help participants reason through tradeoffs without prescribing their choices. Our hypothesis was that these adjustments would lead to clearer demonstration of the relationship between alignment and trust. Lastly, we tested a new method of developing probability distributions from limited amounts of observations, again, to see if this approach would lead to a clearer demonstration of the relationship between alignment and trust. We used the same two KDMAs, VoL and QoL. For each attribute, ADSs were trained to replicate behavior consistent with KDE-modeled human decision profiles. To enhance salience and clarity, each scenario was trimmed to include only the six probes used during the delegation phase, and TRISS scores were added to summarize visible patient vitals. Delegators were shown both the decisions made by the ADS and the unchosen options, giving them context for evaluating trust.

Results (Table 3) showed substantial progress relative to earlier studies. Alignment scores exhibited greater spread, and the behavior of tuned ADSs diverged more clearly from baseline ADSs. For the VoL attribute, alignment showed a stronger relationship with trust, suggesting that the revised probe structure and scenario enhancements were effective. QoL, however, continued to show weak or inconsistent results. Delegators were more likely to trust decision-makers who matched their own KDMA profiles in VoL scenarios, even when they were unaware that the ADS was an artificial agent.

Table 3: Study 3 Results. For the VoL KDMA, alignment between participants and ADS predicts willingness of participant to trust (left) and delegate to (right) the ADS.

Alignment Predicts Trust for VoL, not QoL

KDMA	Study 3
VOL	$p < 0.001, r = 0.27$
QOL	$p = 0.48, r = 0.04$

Alignment Predicts Delegation for VoL but not QoL

KDMA	Study 3
VOL	$\hat{p} = 0.69, p < 0.01$
QOL	$\hat{p} = 0.53, p = 0.39$

Study 3 validated several core assumptions: (1) The use of KDEs, JSD-based scoring, and distributional alignment modeling proved scalable and interpretable in operationally relevant scenarios; (2) TRISS cues improved clarity without constraining judgment, allowing us to use smaller, six-probe scenarios; (3) Alignment does predict delegation and trust – but only for VoL. Study 3 also highlighted challenges. Post-hoc analyses highlighted the need for improved ways of scoring KDMAs, particularly for underperforming KDMAs like QoL. These analyses motivated us to develop a "calibration-based" KDMA measurement approach.

Population Calibration Approach

In complex decision-making domains like medical triage—where there is often no single "correct" answer—aligning AI systems with human values is particularly challenging. Traditional methods that assign fixed scores to decisions based on expert judgments can introduce bias and fail to capture the diversity of human reasoning across different contexts. To address this, we developed a calibration-based approach that estimates a person's decision-making profile from a small set of observed choices. This approach draws on patterns learned from a larger calibration population, who completed both simplified tradeoff tasks (calibration probes) and more complex, realistic decision scenarios (evaluation probes) targeting Key Decision-Making Attributes (KDMAs) such as Value of Life (VoL) and Quality of Life (QoL). From the calibration population's responses, we generated reference distributions for each probe—representing the likelihood of observing particular choices given a specific KDMA value. When a new participant (human or AI) completes the same probes, we compare their response patterns to these reference distributions to infer their KDMA profile.

To assess alignment between human and AI decision-makers, we re-calculated the Study 3 alignment scores using the calibration scores for each KDMA. Compared to fixed scoring methods, the calibration approach yielded more accurate and individualized KDMA profiles, which in turn better predicted participants' trust in and willingness to delegate to AI systems (Table 4). These findings suggest that calibration-based modeling can meaningfully enhance human-AI teaming by grounding alignment in empirically observed decision behavior.

Table 4: Study 3 results recalculated using the calibration approach to measuring KDMAs. All other analyses methods were unchanged from the DRE. Now, for both VoL & QoL KDMAs alignment between participants and ADS predicts willingness of participant to trust (left) and delegate to (right) the ADS.

DISCUSSION: DESIGNING TRUSTWORTHY, ADAPTIVE AI SYSTEMS

Alignment Predicts *Delegation for Both VoL and QoL*

KDMA	Single-number Scoring	Calibration Scoring
VOL	$\hat{p} = 0.69, p < 0.01$	$\hat{p} = 0.71, p < 0.01$
QOL	$\hat{p} = 0.53, p = 0.39$	$\hat{p} = 0.65, p = 0.03$

Alignment Predicts *Trust for Both VoL and QoL*

KDMA	Single-number Scoring	Calibration Scoring
VOL	$p < 0.001, r = 0.27$	$p = 0.000, r = 0.322$
QOL	$p = 0.48, r = 0.04$	$p = 0.000, r = 0.362$

Defining AI Trustworthiness

In high-stakes decision-making environments, trust in AI systems is paramount. This research operationalizes trustworthiness through a measurable alignment score, quantifying the congruence between AI recommendations and human decision-making attributes, known as KDMAs. Our results suggest that alignment correlates with increased user trust and a greater propensity to delegate decisions to AI systems. This alignment-based approach offers a quantifiable metric for trustworthiness, moving beyond traditional reliance on explainability or performance metrics

alone. By focusing on the alignment of AI outputs with human cognitive and ethical frameworks, this method provides a more nuanced understanding of trust in AI systems.

Adaptive AI Design Based on Decision-Maker Attributes

The integration of KDMAs into AI systems facilitates dynamic adaptation to individual user profiles. By modeling attributes such as risk tolerance, value prioritization, and frugality, AI systems can tailor their recommendations to align with the user's decision-making style. This personalization enhances the AI's role as a collaborative partner rather than a mere tool, suggesting that AI systems capable of adapting to user-specific KDMAs would not only increase trust but also improve decision-making efficiency. This adaptive capability is particularly valuable in environments characterized by uncertainty and time pressure, where personalized support can significantly impact outcomes.

System Integration and Deployment

For practical deployment, KDMA-based alignment should be integrated into modular decision-support frameworks. Such integration allows for scalability across various roles, missions, and operational domains. The modular approach supports rapid onboarding of new user roles and facilitates updates to scenario libraries linked to KDMA representations. Implementing this architecture in real-world settings requires systems that can function effectively in dynamic, data-constrained environments, such as military and healthcare contexts. The adaptability and scalability of KDMA-informed AI systems make them well-suited for these applications.

Embedded Ethics and Accountability

Ethical considerations must be embedded throughout the AI development lifecycle. Drawing inspiration from DevSecOps, the integration of Ethical, Legal, and Social Implications (ELSI) principles should occur from the initial design stages. This proactive approach ensures that AI systems are developed with a focus on transparency, fairness, and accountability. Incorporating ELSI principles early in the development process is especially critical in military and clinical decision-support systems, where the consequences of AI recommendations can be profound. By addressing potential biases and ensuring decision traceability from the outset, developers can create AI systems that are both trustworthy and aligned with stakeholder expectations.

In medical triage, these concerns are heightened when providing personalized decision support. While aligning AI with individual KDMA profiles can enhance trust and delegation, it also raises accountability questions: outcomes may diverge from standards if the system skews towards a single or a few decision-maker’s reasoning style. Personalization also risks amplifying individual biases, such as deprioritizing patients with severe disabilities under strong Quality of Life orientations. More broadly, KDMA-based systems balance standardization and personalization. Standards support consistency across providers, while personalization offers cognitive compatibility for individual users. Responsible design requires that KDMA-informed systems complement rather than override doctrinal baselines, ensuring both ethical safeguards and operational trust.

Although our experiments focused on medical triage, these findings support broader principles for AI design in high-stakes, ambiguous domains. First, alignment should be attribute-centered rather than outcome-centered, recognizing that individual reasoning styles influence trust as much as technical accuracy. Second, probabilistic profiles offer a scalable way to capture variability in human judgment across domains, providing richer alignment targets than point estimates. Finally, personalization must be bounded by doctrinal or ethical baselines, offering salient cues without overriding human agency. These principles extend beyond triage to contexts such as C4ISR, Rules of Engagement, and humanitarian assistance, where decisions are complex, contested, and value-laden.

CONCLUSION AND FUTURE WORK

Trust in AI systems is particularly difficult to establish in complex, high-stakes environments—where time pressure, uncertainty, and ethical ambiguity make traditional ground-truth alignment strategies insufficient. Our work begins with the premise that individual differences in reasoning matter, and that these can be captured through KDMAs—stable, psychologically grounded traits that shape how people interpret situations and make tradeoffs. To faithfully model these traits, we represent KDMAs as probability distributions rather than static traits, capturing both

consistency and variability in decision behavior. These profiles can be embedded into AI systems, enabling the AI to mirror human reasoning styles. By aligning system outputs with user KDMA profiles, we provide a foundation for more trustworthy, cognitively compatible decision support. Our experimental results suggest that this alignment meaningfully affects both user trust and willingness to delegate—even when users are unaware the decision-maker is an AI.

This research underscores the potential of KDMA-based modeling in enhancing the trustworthiness of AI systems operating in complex, high-stakes environments. By aligning AI recommendations with individual decision-making attributes, these systems foster greater user trust and facilitate more effective human-AI collaboration. While our research focused on point-of-injury medical triage, the underlying cognitive demands—rapid decision-making under uncertainty, ethical tradeoffs, and mission-critical prioritization—are shared across other military domains. Applications such as C4ISR, HADR, ROE, involve similarly complex environments where decision support must reflect individual reasoning. The KDMA framework's ability to model decision variation and support alignment-based trust evaluation makes it broadly suitable for AI integration in these operational contexts.

Future research directions include:

- **Longitudinal Modeling**: Investigating how decision-maker profiles and trust in AI systems evolve over time
- **Team-Level Alignment**: Exploring the dynamics of AI alignment within teams, considering collective decision-making processes
- **Field Deployment**: Implementing and evaluating KDMA-informed AI systems in live operational settings to assess their impact on decision-making efficacy and trust

Given the scrutiny often associated with medical triage decisions, KDMA-informed AI systems can provide transparent justifications for recommendations, potentially reducing retrospective second-guessing and increasing confidence in frontline decisions. Moreover, the principles and methodologies developed in this research are applicable beyond the medical domain, extending to areas such as ISR, Command and Control (C2), and other complex military decision-making contexts where trust and adaptability are essential.

ACKNOWLEDGEMENTS

The research reported in this document was performed in connection with contract number FA8650-23- C-7315 with U.S. Air Force Materiel Command (USAF/AFMC) and Defense Advanced Research Projects Agency (DARPA). Approved for public release; distribution is unlimited. The views and conclusions contained in this document are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of ACC-APG, USAF/AFMC, DARPA, or the U.S. Government unless so designated by other authorized documents.

REFERENCES

Altay, N., Heaslip, G., Kovács, G., Spens, K., Tatham, P., & Vaillancourt, A. (2024). Innovation in humanitarian logistics and supply chain management: a systematic review. Annals of Operations Research, 335(3), 965-987.

Appelt, K. C., Milch, K. F., Handgraaf, M. J. J., & Weber, E. U. (2011). The Decision Making Individual Differences Inventory and guidelines for the study of individual differences in judgment and decision-making research. Judgment and Decision Making, 6(3), 252–262.

Arend, R. J. (2020). Strategic decision-making under ambiguity: A new problem space and a proposed optimization approach. Business Research, 13(3), 1231–1251.

Baeza, V. M., Parada, R., Salor, L. C., & Monzo, C. (2025). AI-Driven Tactical Communications and Networking for Defense: A Survey and Emerging Trends (arXiv:2504.05071). arXiv.

Benz, N. L. C., & Rodriguez, M. G. (2025). Human-Alignment Influences the Utility of AI-assisted Decision Making (arXiv:2501.14035).

Cojocar, W.J. (2011). Adaptive Leadership in the Military Decision Making Process. Mil. Rev. Nov-Dec, P. 29-34.

Hu, B., Ray, B., Leung, A., Summerville, A., Joy, D., Funk, C., & Basharat, A. (2024). Language Models are Alignable Decision-Makers: Dataset and Application to the Medical Triage Domain (arXiv:2406.06435).

Hu, M., Don, H. J., & Worthy, D. A. (2025). Distributional dual-process model predicts strategic shifts in decision-making under uncertainty. Communications Psychology, 3, 61.

Huelss, H. (2024). Transcending the fog of war? US military 'AI', vision, and the emergent post-scopic regime. European Journal of International Security, 1–21.

Korem, N., Duek, O., Jia, R., Wertheimer, E., Metviner, S., Grubb, M., & Levy, I. (2025). Modeling decision-making under uncertainty with qualitative outcomes. PLOS Computational Biology, 21(3), e1012440.

Lin, J. (1991). Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory, 37(1), 145–151.

Lopez, J., Textor, C., Lancaster, C., Schelble, B., Freeman, G., Zhang, R., McNeese, N., & Pak, R. (2024). The complex relationship of AI ethics and trust in human–AI teaming: Insights from advanced real-world subject matter experts. AI and Ethics, 4(4), 1213–1233.

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.

Mok, A. (2025). How AI and robotics can help prevent breakdowns in factories—And save manufacturers big bucks. Business Insider. Retrieved May 14, 2025, from https://www.businessinsider.com/artificial-intelligence-robotics-predictive-maintenance-manufacturing-factory-solutions-2025-5.

Molineaux, M., Weber, R. O., Floyd, M. W., Menager, D., Larue, O., Addison, U., Kulhanek, R., Reifsnyder, N., Rauch, C., Mainali, M., Sen, A., Goel, P., Karneeb, J., Turner, J., & Meyer, J. (2024). Aligning to Human Decision-Makers in Military Medical Triage. In J. A. Recio-Garcia, M. G. Orozco-del-Castillo, & D. Bridge (Eds.), Case-Based Reasoning Research and Development (pp. 371–387). Springer Nature Switzerland.

Oreg, S., & Bayazit, M. (2009). Prone to Bias: Development of a Bias Taxonomy from an Individual Differences Perspective. Review of General Psychology, 13(3), 175–193.

Sarkar, U. E. (2025). Evaluating alignment in large language models: A review of methodologies. AI and Ethics.

Schluter, P. J., Nathens, A., Neal, M. L., Goble, S., Cameron, C. M., Davey, T. M., & McClure, R. J. (2010). Trauma and Injury Severity Score (TRISS) Coefficients 2009 Revision. Journal of Trauma: Injury, Infection & Critical Care, 68(4), 761–770.

Shattuck, L. G., Miller, N. L., & Kemmerer, K. E. (2009). Tactical Decision Making under Conditions of Uncertainty: An Empirical Study. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 53(4), 242–246.

Shortland, N., Alison, L., & and 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.

Shortland, N., & and Alison, L. (2020). Colliding sacred values: A psychological theory of least-worst option selection. Thinking & Reasoning, 26(1), 118–139.

Shortland, N. D., Alison, L. J., & Moran, J. M. (2019). Conflict: How soldiers make impossible decisions. Oxford University Press.

Silverman, B.W. (1998). Density Estimation for Statistics and Data Analysis (1st ed.). Routledge, New York.

Tejeiro, R., Alison, L., González, J. L., & Shortland, N. (2023). 'Let's be careful out there': Maximization and core values predict action time in police decision making. Personality and Individual Differences, 215, 112398.

Thompson, M. M., Hendriks, T., & Blais, A.R. (2018). Military Ethical Decision Making: The Effects of Option Choice and Perspective Taking on Moral Decision-Making Processes and Intentions. Ethics & Behavior, 28(7), 578–596.

Vestner, T. (2024). From strategy to orders: Preparing and conducting military operations with artificial intelligence. In R. Geib & H. Lahmann (Eds.), Research handbook on warfare and artificial intelligence. Edward Elgar Publishing. Cheltenham UK.

Zamanan, H. M. S. A., Al-Yami, M. Y. M., Dowais, R. M. S. A., Haydar, M. A. M. A., Hokash, N. S. A. A., & Alabbas, H. M. (2022). Navigating Chaos: A Critical analysis of Decision-Making in Emergency Medicine. Journal of Population Therapeutics and Clinical Pharmacology, 29(03), 1568-1578.