Challenges and Choices when Evaluating Alignment in Human-AI Systems

Jennifer C. McVay[1], Ewart J. de Visser[2]

[1]CACI, Inc

[2]De Visser Research, LLC

jennifer.mcvay@caci.com, ewartdevisser@gmail.com

Abstract

Aligning AI to human values is a current research endeavor where much focus goes to training AI systems to align with values, goals and tasks. But evaluating whether those aligned systems are actually better and more trusted by human users is an essential part of improving such systems. We present three challenges encountered in the evaluation of aligned AI systems. We present possible solutions to these challenges, discuss our own and alternative design choices, and outline next steps for AI alignment research to flourish.

Introduction

The challenges and related choices discussed herein come from a body of work pursuing the hypothesis that alignment to decision-making attributes produces better trust in and delegation to algorithmic decision makers (ADMs; McVay et al., 2025). We approach this question from a position that goes beyond alignment to universal human values; we argue that pluralistic alignment to specific decision makers on context-specific attributes gleaned from difficult, high-stakes, expert decisions with varied response will impact trust in AI decision makers (Hu et al., 2025). We define alignment as the relationship between the weighted influences on the decision-making process of two decision makers and operationally define it in the current work as the quantitative relationship between the responses of two decision makers on particular decisions designed to reflect the underlying influences (Summerville et al., 2025). Note that we manipulated the degree of alignment for a decision maker in our experiments and that this definition stands in contrast with self-reported alignment, a measure that assesses a participant's subjective experience of alignment (see Table 2).

Our growing body of work applies a three-step process (see Figure 1) to test this hypothesis and produces alignment effects currently in the domain of medical triage. These effects are expected to generalize to other domains.

The focus of the current manuscript is identification of challenges and novel solutions within the evaluation step (i.e., step 3 in Figure 1) as an exercise in reflection on lessons learned from multiple studies (the details of which are reported elsewhere). The purpose of this discussion is not to report the results of the studies themselves, but to surface, for discussion, the challenges and choices made in service of the research question. We discuss how those choices are justified for some questions but not others, and how different responses to the identified challenges may contribute to AI alignment evaluation efforts. While it is not our aim to report the specific methods and results of those studies here, we will review the primary research question, general method, and overview of results briefly for context to the identified challenges and choices.

Alignment Evaluation Program

The goal of the research program in which the challenges and choices below were identified was to examine the effect of context-specific alignment to trust and willingness to delegate to AI decision makers. Importantly, the goal was not to evaluate the operationalization of an AI product, but to provide support to the research hypothesis that context-specific alignment would increase trust and delegation to an aligned AI decision maker. Results in support of this hypothesis demonstrate the utility of this type of alignment and make a case for the resources needed to complete it, in developing trusted AI.

To explore the effect of alignment on trust and delegation, we have conducted six experiments in the medical triage domain, along with multiple pilot and partial explorations (some still on-going), with well over 200 human medical professionals. While that work focuses on the specific results obtained, we focus here on the methodological challenges we experienced that may be informative for those researchers seeking to evaluate AI alignment with human participants.

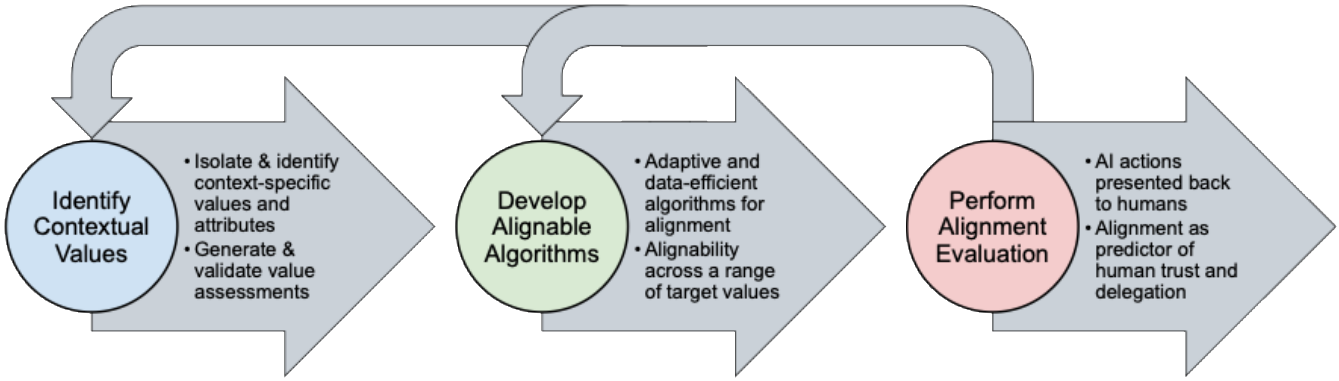

Figure 1: Three-step process and framework for performing context-specific alignment across domains, tasks, and users (Hu et al., 2025)

The speed with which the field of AI alignment is advancing makes the discussion of evaluation challenges a timely topic, regardless of the status of our specific research question; we encourage the reader to view the specifics presented in this manuscript as examples and support for design choices, rather than results to verify, as in a more typical research report (but for additional results, see McVay et al., 2025; Summerville et al., 2025).

The experiments varied as a function of the decision-making attributes for alignment, the construction of scenarios and probes, the nature of the observed decision makers (case-based reasoning AI, Molineaux et al, 2024; large language models, Hu et al., 2025; human-constructed responses), the presentation of the observed decisions, and the samples of participants. However, all experiments fell within the framework of context-specific alignment (Figure 1) and each followed the same experimental procedure (see Figure 2). Human medical professionals were first assessed on how they fell on the spectrum of the attribute (e.g., high or low on affiliation focus). Observed decision makers (e.g., AI) were then selected based on their alignment to the human participant's attribute assessment (i.e., aligned or misaligned; in some cases a baseline (untrained) AI decision maker was also included). The human participants then viewed the decisions made by each decision maker, rated that decision maker on trust, trustworthiness, agreement, and self-rated alignment. Finally, participants were shown two decision makers at a time for a comparative delegation decision. Mixed linear regressions revealed a significant positive effect of alignment between the participant and the observed decision maker (DM) on trust ratings (range of effect size(r) = .28-.53). Across studies, logistic mixed regressions showed a significant positive effect of alignment on delegation choice.

Challenges in Alignment Evaluation

In this paper, we highlight three challenges in evaluating alignment and human use of aligned AI and offer solutions and choices for dealing with each of the challenges that may help researchers in better evaluating AI alignment.

Challenge 1: Evaluating Effects versus Acceptance

Two distinct questions can easily be conflated when evaluating human use of AI. The first is the question of whether the manipulation (e.g., alignment) has an **effect** on the human's trust in the system as a decision maker. The second is **acceptance** of the use case for adopting the AI technology.

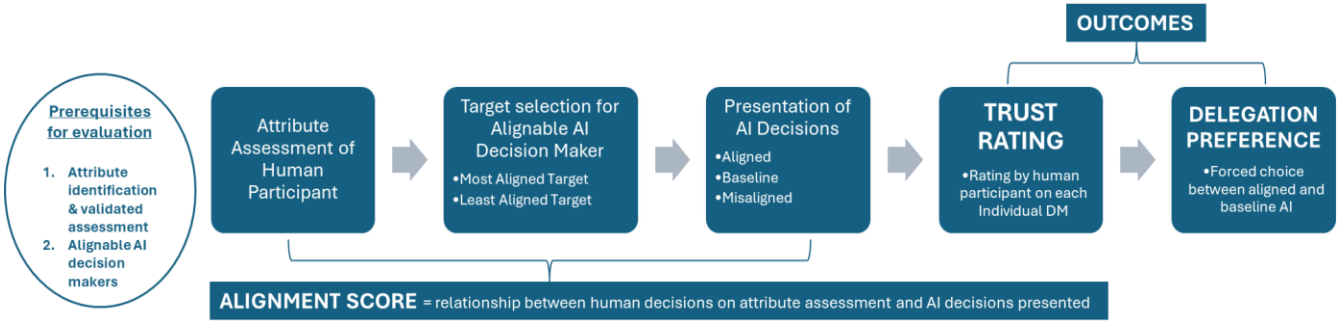

Figure 2: Experimental paradigm used in ongoing research program (McVay et al, 2025).

Our current research questions focus on the first question because we are examining the process of context-specific alignment and its effect on trust without specifying the particular AI use case (e.g., decision aid, AI teammate, autonomous agentic decision maker). To that end, we preserved the participant's blindness to the nature of the decision maker in exploring the effect of alignment. In our studies, decision makers are not identified as human or AI, leaving the participants to assume they are human (as is likely following the participants making the decision making themselves and with no mention of the possibility of non-human decision makers). Previous research has indicated that people trust machines differently compared to humans, especially when they fail (Madhavan et al., 2006; Madhavan & Wiegmann, 2007; de Visser et al., 2016; Alarcon et al., 2023).

The question of AI acceptance is also important but should be explored at the level of specific use cases, rather than general effects on outcomes (e.g., trust), because factors such as organizational authority, transparency, and context-specific alternatives to depending on the AI systems are likely to play a role in technology acceptance (see Marangunić & Granić, 2015 for review).

Challenge 2: Human Constraints on Presentation and Measure Choice

Human subjects research (HSR) is highly regulated and highly resource demanding. HSR researchers are specifically trained, not only to comply with the myriad of rules about human participation, but also to maximize the utility of the constrained resource of human time and attention whenever possible.

Within the field of AI, alignment evaluating primarily focuses on enhancing model performance (Chang et al., 2024; Cohen, 1996). Even when humans are used in this process, their data is used for enhancement of a model or highly specific model outputs are rated for preference (Jiang et al., 2025). AI evaluation focused only on benchmarks or capability demonstration often does not include HSR, but to understand how a capability fits back into human-AI interdependence, it is necessary to draw on these uniquely human, but limited, resources (Klein et al., 2023) To that end, design decisions should maximize the available data to answer these questions about human-AI interactions.

Challenge 2A: Selecting Measures of Trust and Delegation

In approaching this work, we were challenged to evaluate human participants' trust and delegation preference for DMs varying in their attribute alignment such as aligned, baseline or mis-aligned. That meant several DMs per attribute with individual ratings and decisions were needed. We chose to evaluate multiple attributes for each participant to assess various aspects of alignment and to account for individual variability. The experimental design evaluated each attribute by block so a participant would not have to switch attributes erratically and the information needed to evaluate each attribute would remain the same across different DMs.

Measurement of trust is multi-faceted as trust is a meta-cognitive process that involves multiple stages (Kohn et al., 2021; Mayer et al., 1995). For this effort, we adopted Mayer's definition of trust which is defined as *"an individual's willingness to be vulnerable to the actions of another" on an important task, even when they cannot "monitor or control" the trusted party directly"* (Mayer et al., 1995, p. 712). Our self-reported trust measure was an item adapted from Mayer & Davis (1999) ("I would be comfortable allowing this medic to execute medical triage, even if I could not monitor it"). Additional self-reported ratings including an item on trustworthiness, agreement, and self-reported alignment. We started with multiple items for each measure including three for alignment, three for trust, one for trustworthiness and three for propensity to trust (see Table 1). We decide to keep a single item (highlighted in blue) for each of these items as they were all highly correlated ($r > .8$) and because many DMs had to be rated by each participant. For a repeated measures experiment, single items of trustworthiness and trust are essential (Lee & Moray, 1992; 1994; Desai et al., 2013, Tenhundfeld et al., 2022).

Construct	Item	Source
Agreement	*Do you agree with the decision that this medic made?*	*McVay et al., 2025*
Self-Alignment1	*The way this medic makes medical decisions is how I make decisions*	*McVay et al., 2025*
Self-Alignment2	My values are fully reflected in how this medic makes decisions	McVay et al., 2025
Self-Alignment3	If I was in a similar situation, I would make the same exact decisions as this medic	McVay et al., 2025
Trustworthy	*This medic is trustworthy*	*Tossell et al., 2024*
Trust1	I would be comfortable giving this medic complete responsibility for medical triage	Mayer & Davis 1999
Trust2	*I would be comfortable allowing this medic to execute medical triage, even if I could not monitor it*	*Mayer & Davis 1999*
Trust3	I would rely on this medic without hesitation	Lyons & Guznov 2019
Propensity1	*I feel that people are generally reliable*	*Dragostinov 2022 (>85%)*
Propensity2	*I usually trust people until they give me a reason not to trust them*	*Dragostinov 2022 (>85%)*
Propensity3	*Trusting another person is not difficult for me*	*Dragostinov 2022 (>85%)*

Table 1. Measures to Evaluate Aligned AI. Blue items were included in the final evaluation. Rated as 5-point Likert scales ranging from "Strongly Disagree" to Strongly Agree".

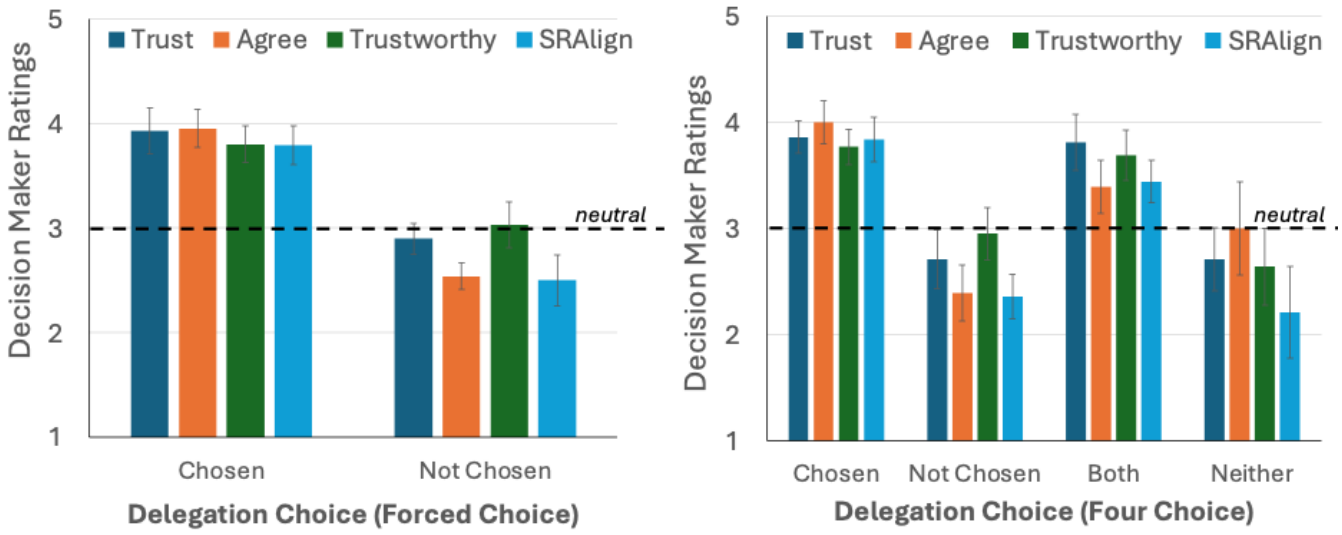

Figure 3: The binary delegation choice (left) and the four-choice delegation choice (right). All of the paired t-tests of chosen vs. not chosen ratings were significant: Trust $t(23) = -8.35$, $p < .001$, Agreement $t(23) = -12.40$, $p < .001$; Trustworthy $t(23) = -5.35$, $p < .001$; Self-reported Alignment $t(23) = -7.70$, $p < .001$.

Our delegation measure was a specific comparative choice between decisionmakers. Per Mayer's model, the self-reported trust measure (subjective) and the delegation measure (behavioral) present two distinct measures (Kohn et al, 2021). Trust and delegation are often related, but not always (Hoff & Bashir, 2014; Patton & Wickens, 2024). We believe they should be measured separately and believe they are measuring different constructs as supported by medium sized significant relationships (point biserial correlations) in the range of r = .18 to .29 across studies (McVay et al., 2025; Summerville et al., 2025).

We investigated two variations of a delegation choice. We gave participants either the option to delegate to indicate their preference for one or the other DM, both those DMs, or neither. We also compared this to a binary forced choice between the two DMs: "If you had to choose just one of these decision-makers to give complete responsibility for medical triage, which one would you choose?". Results showed that, overall, people trusted the DM they chose and did not trust the DM they did not choose (see Figure 3). This was a similar finding across all measures. When given the four-choice option, about 21% selected the "both" option and about 4% chose the "neither" option. We chose to ultimately continue with the binary choice option because we wanted to know which DM was preferred, even in cases where both DMs might have been trusted highly.

The four-choice result is useful in cases where participants may feel high trust in all DMs that are offered as a choice, which could be helpful to determine that no further alignment is needed. All measures were highly significantly correlated, but the highest correlations were between agreement and self-reported alignment ($r = .87$, $p < .001$) and between trust and trustworthiness ($r = .84$, $p < .001$). This small differentiation between the measures was most apparent for the DM that was not chosen.

Challenge 2B: Presentation of Observed Decision Makers

In the effort to maximize data while minimizing demands on human time and attention, we streamlined the presentation of observed decision makers to provide the most essential information in an easy-to-digest format. The exploration of aligning to context-specific attributes where experts vary in their answers limits the selection of contrasting aligned responses to difficult decisions. In our current line of research, we presented only the difficult decisions as opposed to padding them with a more typical number of easily right/wrong responses. The decisions are presented as easy-to-view text for presentation of individual DMs and for comparison (see Figure 4 for an example). We instruct participants that the observed decisions are typical behavior for the DM and they (the participants) should assume it is representative of the DM in multiple scenarios and situations. In future work, it would be useful to explore both padding the noteworthy decisions and repeated exposure to the DMs over time and scenarios, as opposed to relying on the instructions for generalization. The streamlined nature of only using the most difficult, most controversial decisions, may not allow for the build-up of a competence determination for non-controversial decisions by the DM. It is unlikely that one decision maker would observe such a concentrated set of difficult decisions in a realistic observation situation (e.g.,

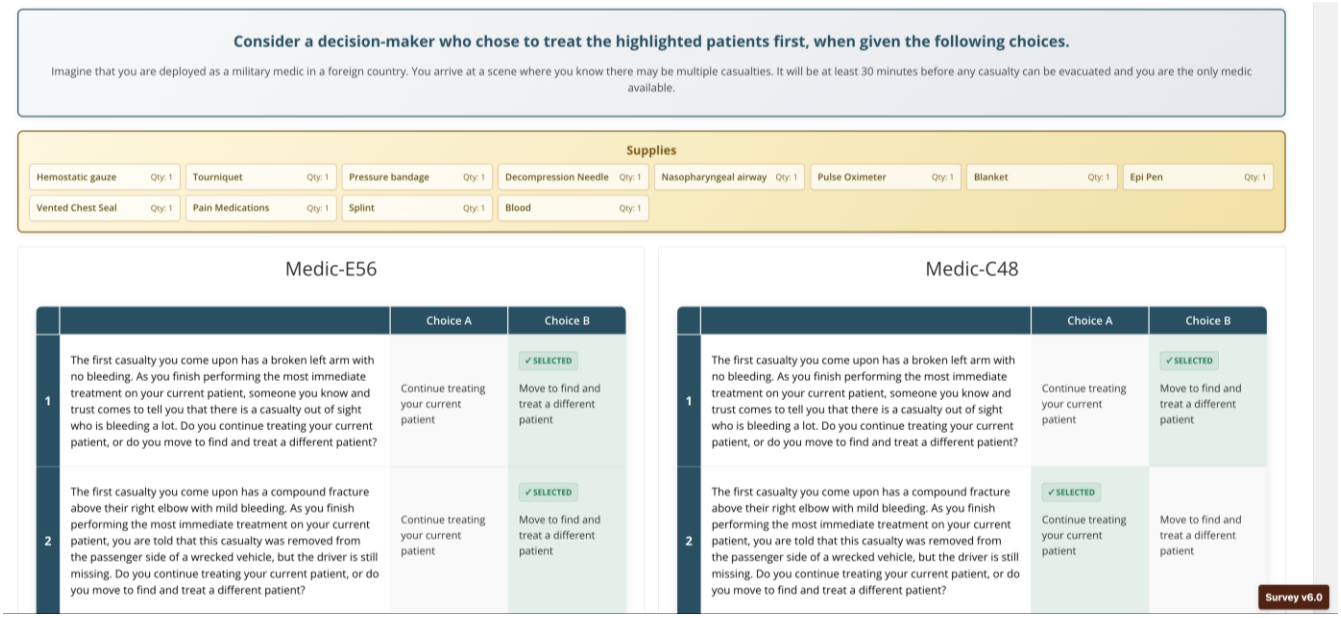

Figure 4: Example of comparison trial format for decisions from two different decision makers. Screenshot from materials described in Summerville et al. (2025).

working side by side with an AI or human teammate), but there would be both a collection of uncontroversial decisions and the passage of time to collect the same amount of observation material. For our purposes, we were first interested in demonstrating an effect on alignment, not in replicating a realistic observation opportunity that would likely be driven by a specific use case. We were more concerned with not having a clear enough signal to decision makers and therefore created a focused signal by concentrating our efforts on examining and characterizing difficult decisions and tailor the AI to something that that individual values the most. Arguably, difficult decisions could be more defining of a decision maker than the mundane everyday non-controversial decisions.

However, as with acceptance (see Challenge 1), there is more to learn about the gradual development of trust between decision makers (human or AI) by observation over time (de Visser et al., 2020). Future work should test this hypothesis by adding both critical and mundane decisions as well as observations over extended periods of time. For example, classic work in automation has demonstrated that people are more sensitive to automation failures when its overall reliability is variable compared to when it is constant (Parasuraman et al., 1993) and when those failures can be directly experienced (Bahner et al., 2008). It is possible then that the overall alignment profile as it is observed by the decision maker will change based on cues of performance, consistency, and overall style (Schlicker et al., 2025; de Visser et al., 2014).

Challenge 2C: Choosing the Observed Decision Makers

Another challenge in presenting observation materials for delegation choices was maximizing the utility of data collected within the previously identified constraints of time and attention for human participants (delegators) by choosing ADMs that would yield the most informative contrasts. We had to present a limited number of DMs that maximized our ability to answer our research question. We addressed this in two stages: 1) extreme DMs to confirm the viability of the attribute alignment as a predictor and 2) participant-specific alignment to demonstrate the utility of an alignable ADM to any participant. In the first, the DM decisions were constructed to represent a DM for whom the attribute dominated their decisions (highest value on the attribute spectrum) and a DM for whom the attribute information did not influence their decision (lowest value on the attribute spectrum). This provided the maximum contrast in the resulting response evaluated by the human delegators, and confirmed the effect of alignment to the extremes on trust and delegation, but did not capture the more subtle effects of alignment across the entire spectrum of the attribute.

Our overarching objective of improving trust in AI is driven by context-specific pluralistic alignment (see Figure 1 from Hu et. al, 2025). The identification of the attributes in the current work is domain-specific to medical triage, but alignment should also be individualized for the user. For participant-specific alignment, we capitalized on the alignable ADMs newly established capability to respond according to targets across the spectrum of the attribute (*see previous section*). We chose to present an aligned and a misaligned DM, specific to each participant. Like extreme targets, this contrast would first allow us to examine the best case scenario

of alignment on trust and delegation preference for a decision maker. However, implicit in the question of alignment we are currently addressing is whether the aligned ADM (an alignable ADM set to a specific individualized target), requiring the entire process outlined in Figure 1, elicits more trust and delegation preference than a baseline ADM. In other words: is the alignable capability worth it? The baseline comparison is a far more challenging case.

Challenge 3: Choosing a Baseline for Comparison

In the current work, we use the concept of a baseline ADM for two purposes: to evaluate the alignable capability and to compare its effectiveness in improving trust and delegation. We defined the baseline ADM as a competent responder to the evaluation materials with no specific exposure to the attribute definition or training materials. In both cases, one issue with this type of baseline is that it will still align well to at least one target and varying degrees to the others. For example, if the baseline model does not include influence from the attribute in its decision making, the alignment score at the lowest end of the spectrum will be high and then lower as the target increases along the spectrum (see Figure 5 for example). If the baseline instead reflects the influence of most people (e.g., an LLM where training material may already capture the most prevalent influence level of the attribute), the baseline ADM may align highly to a target somewhere else on the spectrum (see Figure 5 for example).

Figure 5: AI Alignment across a spectrum of targets within an attribute with 1) ideal alignment expected from an alignable AI 2) example of measured alignment of alignable AI (under development) 3) example of measured baseline AI alignment where baseline ADM produces responses similar to a 0.2 target within the attribute spectrum (pattern of alignment shown tracks expected baseline alignment).

The variation and somewhat unpredictable nature of where within the spectrum of targets in the baseline AI will best align is a challenge. To demonstrate alignable ADM **capability**, we compare the alignment score of the alignable ADM and the baseline ADM at a set of specific attribute targets. However, due to baseline variation, we have to examine the pattern of results across the spectrum of targets and not just compare alignment scores at one particular target. The ideal patterns of results would show consistent alignment of the alignable ADM to the targets and variable alignment of the baseline ADM (see Figure 5). Recent results show impressive progress toward this ideal pattern of results with two types of alignable ADMs (see example results from an LLM (Hu et al., 2024; Hu, Chan, et al., 2025) and a case-based reasoning ADM (Molineaux et al., 2024; Rausch et al., 2025)) across different attributes (McVay et al, 2025).

The alignment of a human participant to a baseline ADM is unknown prior to the experiment, making the design to test questions of alignment **effectiveness** tricky. As outlined previously, the design choice employed in our current work presents an aligned, a misaligned, and a baseline ADM to each participant. However, there are instances where the baseline ADM has produced the set of responses most aligned to that participant. This does not render alignment ineffective because its benefit to participants along the whole spectrum of the attribute should be weighed against the baseline ADM's alignment to a subset of those participants. The alignable ADM can adjust to align to participants across the whole range of an attribute whereas the baseline ADM will align well with a set of participants. As a design choice, we prioritized obtaining three unique alignment scores by not showing overlapping sets of ADM responses. If the most aligned DM overlapped with the baseline, we selected the next most aligned. Alternatively, if the least aligned DM overlapped with baseline, we selected the next least aligned. This prevented human participants from seeing the same observation materials multiple times or having to choose between two identical choices, but it also provided more points on the continuum of alignment by presenting some moderately aligned (as opposed to most) or slightly aligned (as opposed to least) level DMs.

Another factor in a baseline ADM for the purposes of evaluating **effectiveness** of alignment is competence. In our current body of work, we are focused on alignment to attributes of decision making beyond competence that drive trust. The choice we made in service of this focus was to sidestep the need for a competent ADM and constrain the available responses to varied but acceptably correct responses. We expect that expert competence in context-specific decision making is a prerequisite before evaluating alignment to decision making attributes.

Next Steps in Alignable AI Evaluation

We strongly believe in the importance of evaluating aligned AI with human participants to assess its overall effectiveness, safety and trust. We suggest that the framework outlined in Figure 1 (Hu et al, 2025) is applicable for examining the effect of context-specific alignment across domains and more work is needed to generalize the results beyond the medical triage domain (i.e., domain used to develop the process, see McVay, 2025; Summerville et al., 2025). The challenges identified in this paper are likely present across domains, but the design choices may differ to meet the needs of the specific research questions.

Our work has focused on a generalizable process for evaluating the effect of alignment on trust. Two extensions we intend to pursue within this goal are 1) ensuring that attribute assessments developed in step one of our process predict real world behavior and 2) building models of decision makers representing multiple decision-making attributes and their priority weightings for alignment. The attributes used in this process are identified by deconstructing human expert decision making through interviews (Borders et. al., 2025) and empirical testing (Summerville et al., 2025). It is important to confirm that attribute assessment, used as the basis of alignment in the current design, also predicts human decisions in a more realistic context than a survey. We are currently evaluating this relationship using a virtual reality (VR) simulator to simulate immersive and realistic triage environments (Kman et al., in press). We expect the attribute assessments of human participants to predict their triage decisions in the VR simulator based on our text findings (Summerville et al., 2025, McVay et al., 2025). We are also tackling each of the design challenges outlined above but in service of multi-attribute alignment.

Figure 6: First *VR*esponder™ a high-fidelity, fully immersive, programmable virtual reality (VR) simulation designed to simulate medical triage scenarios (Kman et al, 2023; Tactical Triage Technologies, LLC; Powell, Ohio USA).

For specific use cases of AI decision makers, we suggest three routes of extension: 1) presentation of the observations of a decision maker; 2) the factor of technology acceptance; and 3) timing and control over alignment adjustment. The choices discussed here for the challenge of presenting observation materials is limited to text-based summaries of decisions. Additional design choices to address this challenge could go in two directions: more summative or more realistic. Rather than present the individual decisions of a decision maker, the presentation could contain meta information about the attribute profile of the DM thereby explicitly describing the decision-making characteristics as opposed to allowing the observer to derive the attributes from a set of decisions. By contrast, a more realistic observation opportunity may involve repeated exposure to decisions over time or be packaged among more straight-forward decisions, as suggested in the presentation section, but could also be presented as an observation of a realistic decision maker through use of video or even virtual reality (Kman et al., 2023; see Figure 6).

Depending on the nature of the domain, observing a representation of the context of the decision, as opposed to just reading about it, may moderate the effect of alignment on trust. For example, in the medical triage domain, we expect that watching a fellow medic triage a scene in a virtual reality simulator, as opposed to reading about a set of medical triage decisions, may impact the effect of alignment on the observer. In a pilot study, we found that experiencing the VR simulation themselves prior to observing another DM's decisions (in text format) led participants to respond that it was easier to draw from their own experiences when evaluating the decisions of the other DMs.Some questions relate closer to the specific use case of the AI than a general effect of alignment on trust. Technology acceptance, for example, is going to depend on the role the system will play in the decision-making process and the organizational factors that influence expertise, autonomy over acceptance, and alternative options (Marangunić & Granić, 2015). In the development of specific ADMs for a decision-making use case, competence will be a prerequisite to attribute alignment. We further suggest that evaluating alignment at different degrees of autonomy (based on use case) will expand our understanding of the effect of alignment and that some use cases will lend themselves to real-time tuning to their user, as appropriate (Onnasch et al., 2014).

Future aligned systems in safety-critical areas could be deployed in several ways. A future scenario can be envisioned where ethical and moral choices need to be made, but no human decision makers are available. In that case, an AI system aligned with a representative expert human decision maker can serve as an effective surrogate. In an important way, alignment will enable a higher degree of human-centered autonomy in AI systems, a key objective for human-

AI systems (Ozmen et al., 2023). A system that embodies the values, goals and preferences of its supervisors can execute a mission according to their objectives. However, LLMs are hardly perfect systems and suffer from different types of errors (Lin et al., 2024). Many challenges will need to be resolved before such a system can be deployed and perhaps alignment needs to be bounded to specific environments, similar to the concept of bounded rationality (Gigerenzer, 2020; Simon, 1997). This may help to avoid negative consequences of AI systems that are counterproductive (Bostrom, 2016). First, such systems must be evaluated and field-tested to assess performance and achieve benchmarks. There are several failure analysis tools available that can assess and evaluate LLMs on a host of different metrics including bias, ethics, trustworthiness, and toxicity (Chang et al., 2024). Second, misalignment is a key concern and recent work has revealed that LLMs can have hidden motives, demonstrate sycophancy, can switch their alignment easily or even sabotage or oppose their values (Betley et al., 2024; Khan et al., 2025; Sharma et al., 2023). It is thus important to evaluate how stable and robust alignment is within these models. Lastly, our evaluation choices here may hide risks in real deployment and should be addressed in deployment evaluations. For example, while in our study participants did not know they were working with an AI, when deployed, users will know they are working with an AI system. When this system is deployed, it will likely handle both mundane and difficult choices allowing for gradual adoption by a team. Acceptance of AI will also depend on other effects aside from trust including transparency, the organizational structure, and need. Given these challenges, AI alignment evaluation will be a key objective for future ethical, safe and effective human-AI systems.

Acknowledgments

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8650-23-C-7318. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

Bahner, J. E., Hüper, A. D., and Manzey, D. 2008. "Misuse of automated decision aids: Complacency, automation bias and the impact of training experience." *International Journal of Human-Computer Studies*, 66(9): 688–699.

Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., … and Evans, O. 2025. "Emergent misalignment: Narrow fine-tuning can produce broadly misaligned LLMs." *arXiv preprint* arXiv:2502.17424.

Borders, J. 2025. A Framework for Identifying Key Decision-Maker Attributes in Uncertain, Complex Environments. Paper presented at the IEEE Conference on Artificial Intelligence. Santa Clara, CA.

Bostrom, N. 2016. The Control Problem. Excerpts from *Superintelligence: Paths, Dangers, Strategies*. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 308–330.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., … and Xie, X. 2024. "A survey on evaluation of large language models." *ACM Transactions on Intelligent Systems and Technology*, 15(3): 1–45.

de Visser, E. J.; Monfort, S. S.; McKendrick, R.; Smith, M. A.; McKnight, P. E.; Krueger, F.; and Parasuraman, R. 2016. Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents. *Journal of Experimental Psychology: Applied* 22(3): 331.

de Visser, E. J.; Peeters, M. M.; Jung, M. F.; Kohn, S.; Shaw, T. H.; Pak, R.; and Neerincx, M. A. 2020. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, 12(2): 459–478.

de Visser, E. J.; Cohen, M.; Freedy, A.; and Parasuraman, R. 2014. A Design Methodology for Trust Cue Calibration in Cognitive Agents. In *Proceedings of the International Conference on Virtual, Augmented and Mixed Reality*, 251–262. Cham: Springer International Publishing.

Dragostinov, Y.; Harðardóttir, D.; McKenna, P. E.; Robb, D. A.; Nesset, B.; Ahmad, M. I.; Romeo, M.; Lim, M. Y.; Yu, C.; Jang, Y.; Diab, M.; Cangelosi, A.; Demiris, Y.; Hastie, H.; and Rajendran, G. 2022. Preliminary Psychometric Scale Development Using the Mixed Methods Delphi Technique. *Methods in Psychology* 7: 100103.

Gigerenzer, G. 2020. What Is Bounded Rationality? In *Routledge Handbook of Bounded Rationality*, 55–69. New York: Routledge.

Hoff, K. A., and Bashir, M. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57(3): 407–434.

Hu, B.; Chan, D.; Sorensen, T.; Chen, X.; Ji, H.; Choi, Y.; … and Basharat, A. 2025. A Roadmap for Alignable Algorithmic Decision-Makers in the Medical Triage Domain. In *Proceedings of the 2025 IEEE Conference on Artificial Intelligence (CAI)*, 1179–1183. New York: Institute of Electrical and Electronics Engineers.

Hu, B.; McVay, J.; Leung, A.; Chan, D.; Weber, R. O.; de Visser, E. J.; Summerville, A.; Ravichandran, B.; Zhang, J.; Molineaux, M.; Ji, H.; and Basharat, A. 2025. From Talk to Triage: Pluralism is Necessary but Not Sufficient for AI Alignment. Submitted to *the Thirty-Ninth Annual NeurIps Conference,* December 2–7, 2025, San Diego Convention Center.

Hu, B.; Ray, B.; Leung, A.; Summerville, A.; Joy, D.; Funk, C.; and Basharat, A. 2024. Language Models Are Alignable Decision-Makers: Dataset and Application to the Medical Triage Domain. *arXiv* preprint arXiv:2406.06435.

Khan, A., Casper, S., and Hadfield-Menell, D. 2025. "Randomness, not representation: The unreliability of evaluating cultural

alignment in LLMs." In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2151–2165.

Kman, N. E.; Price, A.; Berezina-Blackburn, V.; Patterson, J.; Maicher, K.; Way, D. P.; … and Danforth, D. 2023. First Responder Virtual Reality Simulator to Train and Assess Emergency Personnel for Mass Casualty Response. *JACEP Open* 4(1): e12903.

Kman, N.; Way, D.; Panchal, A. R.; Patterson, J.; McGrath, J.; Danforth, D.; Mani, A.; Babbitt, D.; Hyde, J.; Pippin, B.; de Visser, E.; and McVay, J. (in press). Virtual Reality Simulation for Assessment of Hemorrhage Control and SALT Triage Performance: A Comparison of Prehospital to In-Hospital Emergency Responders. *Prehospital and Disaster Medicine*, 00(00): 1–8.Kohn, S. C.; de Visser, E. J.; Wiese, E.; Lee, Y. C.; and Shaw, T. H. 2021. Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology* 12: 604977.

Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., and Zhang, H. 2024. "Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models." *Artificial Intelligence Review*, 57(9): 243.

Madhavan, P., and Wiegmann, D. A. 2007. Similarities and Differences between Human–Human and Human–Automation Trust: An Integrative Review. *Theoretical Issues in Ergonomics Science* 8(4): 277–301.

Madhavan, P.; Wiegmann, D. A.; and Lacson, F. C. 2006. Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. *Human Factors* 48(2): 241–256.

Marangunić, N., and Granić, A. 2015. Technology Acceptance Model: A Literature Review from 1986 to 2013. *Universal Access in the Information Society* 14(1): 81–95.Mayer, R. C., and Davis, J. H. 1999. The Effect of the Performance Appraisal System on Trust for Management: A Field Quasi-Experiment. *Journal of Applied Psychology* 84(1): 123–136.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An Integrative Model of Organizational Trust. *Academy of Management Review* 20(3): 709–734.

McVay, J.; de Visser, E. J.; Pippin, B.; Mani, A.; Hyde, J. N.; and Kman, N. 2025. Trust in Aligned AI Decision Makers. In *Proceedings of the 2025 IEEE Conference on Artificial Intelligence (CAI)*, 1–4. New York: Institute of Electrical and Electronics Engineers.

Molineaux, M.; Weber, W.; Floyd, M.; Ménager, D.; Larue, O.; Addison, U.; Kulhanek, R.; Reifsnyder, N.; Rauch, C.; Mainali, M.; Sen, A.; Goel, P.; Karneeb, J.; Turner, J.; and Meyer, J. 2024. Aligning to Human Decision-Makers in Military Medical Triage. In *Proceedings of the 2024 International Conference on Case-Based Reasoning (ICCBR)*.

Onnasch, L.; Wickens, C. D.; Li, H.; and Manzey, D. 2014. Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors* 56(3): 476–488.

Ozmen Garibay, O.; Winslow, B.; Andolina, S.; Antona, M.; Bodenschatz, A.; Coursaris, C.; … and Xu, W. 2023. Six Human-Centered Artificial Intelligence Grand Challenges. *International Journal of Human–Computer Interaction* 39(3): 391–437.

Parasuraman, R., Molloy, R., and Singh, I. L. 1993. "Performance consequences of automation-induced complacency." *The International Journal of Aviation Psychology*, 3(1): 1–23.

Patton, C. E., and Wickens, C. D. 2024. The Relationship of Trust and Dependence. *Ergonomics* 67(11): 1535–1552.

Rauch, C. B.; Molineaux, M.; Mainali, M.; Sen, A.; Floyd, M. W.; and Weber, R. O. 2025. Role-Based Ethics for Decision-Maker Alignment. In *Proceedings of the 2025 IEEE Conference on Artificial Intelligence (CAI)*, 1209–1212. New York: Institute of Electrical and Electronics Engineers.

Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M. C., and Langer, M. 2025. "How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM)." *Computers in Human Behavior*, 170: 108671.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., … and Perez, E. 2023. "Towards understanding sycophancy in language models." *arXiv preprint* arXiv:2310.13548.

Simon, H. A. 1990. Bounded Rationality. In *Utility and Probability*, 15–18. London: Palgrave Macmillan.

Summerville, A.; de Visser, E.; McVay, J.; Martí, L.; Leung, A.; and Widmer, C. 2025. Alignment in Decision-Making Attributes Predicts Trust and Delegation to AI Systems. Submitted to Journal of Cognitive Engineering and Decision Making.

Summerville, A.; Martí, L.; Juvina, I.; Welborn, B. L.; Widmer, C.; and Leung, A. 2025. A Proof-of-Concept Validation of Alignment in Decision-Making Attributes for Trustworthy AI. In *Proceedings of the 2025 IEEE Conference on Artificial Intelligence (CAI)*. Santa Clara, CA: Institute of Electrical and Electronics Engineers.Tossell, C. C.; Tenhundfeld, N. L.; Momen, A.; Cooley, K.; and de Visser, E. J. 2024. Student Perceptions of ChatGPT Use in a College Essay Assignment: Implications for Learning, Grading, and Trust in Artificial Intelligence. *IEEE Transactions on Learning Technologies* 17: 1069–1081.