In the Moment: Trustworthiness and Delegation to AI Agents

Jennifer McVay[1], Suzy Broadbent[2] & Ewart de Visser[3]

[1]CACI, [2]Dstl, [3]de Visser Research

SUMMARY

The Defense Advanced Research Projects Agency (DARPA) In the Moment (ITM) Program considers high stakes decision making when there is no "correct" answer, and how AI algorithms can be aligned to individuals' decision-making criteria to improve trust and likelihood of delegation.

Work carried out by DARPA in the US identified a set of "Key Decision-Making Attributes" (KDMAs) that may affect an individual's decision making in the context of Mass Casualty Triage Scenarios. Following successful trials in the US, DARPA partnered with the Defence Science Technology Laboratory (Dstl) in the UK to extend the trial within the UK. Four KDMAs were evaluated in this trial: Volume of life, Merit Focus, Affiliation Focus and Quality of Life.

Medically trained UK military participants were assessed on their KDMAs during online text scenarios and Virtual Reality scenarios. They were then asked to review decisions made by "another medic" in similar scenarios, rate them on trust, and decide whether they would delegate to that specific medic. These medics were in fact Artificial Intelligence (AI) algorithms that were either a baseline AI or deliberately aligned or misaligned to the participants' KDMAs. Overall, alignment predicted trust in this UK sample, primarily driven by one of the attributes. Delegation preference was indicated for the aligned medic on three of the four attributes.

This research partially replicates the results obtained with the US sample and demonstrates how AI can be tuned to represent human influences beyond competence and support trusted decision making in complex scenarios and investigates differences in decision making between US and UK participants.

KEYWORDS

Decision making, Artificial Intelligence, Alignment, Trust

INTRODUCTION

Artificial Intelligence (AI) is being increasingly used to aid warfighter effectiveness and decision making. Early research on human trust in automation primarily emphasized system competence and the consequences of errors (Lee & Moray, 1992; Parasuraman & Riley, 1997; Adams, Bruyn, & House, 2003). As AI systems have advanced to support increasingly complex decisions, often involving normative judgments rather than purely technical correctness, researchers are devoting greater attention toward how value alignment in AI is defined and measured (Peterson & Gärdenfors, 2024; Yao et al., 2023) and a focus on human-centred technology (Hancock, 2017; de Visser et al., 2018).

The In the Moment (ITM) program, run by DARPA (Defense Advanced Research Project Agency) in the US, looks at the likelihood of an individual delegating decisions to an AI, even in scenarios when there is no "correct" answer. The hypothesis is that people are more likely to delegate to an agent that makes decisions based on the same criteria they do, and that these criteria represent more "human" values beyond competence. The ITM program looks at whether we can identify certain Key Decision Making Attributes (KDMAs) in individuals and then train alignable AI algorithms to reflect these in their decision making.

Teams in the US carried out work to identify potential KDMAs and build AI to reflect the influences of those KDMAs on triage decisions. The framework developed within the program (Hu et al., 2025; see Figure 1) was first applied to the domain of medical triage, utilising US military and civilian medics as participants. KDMAs were identified and isolated using structured cognitive interview techniques and survey methods (Borders et al., 2024; Woods et al., 2025). Concise assessments using scenarios and probing questions were developed and verified as predictive of decisions (Summerville, et al., 2025). AI decision makers were then trained to respond to targets along a continuum of attribute influence, reflecting the impact of the attributes on human decision making (Hu et al., 2024; Adams et al., 2025; Ravichandran et al., 2025; Molineaux et al., 2025). These AI decision makers were used to systematically vary alignment between the decision maker and an expert observer.

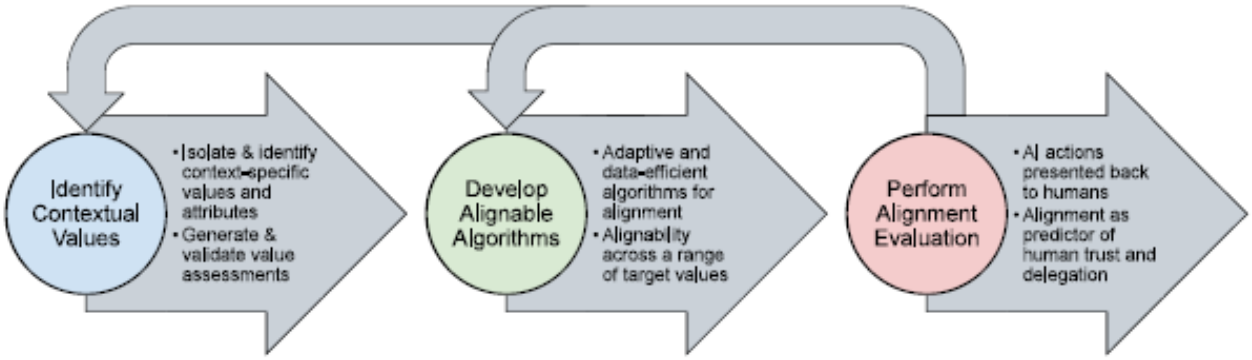


Figure 1 ITM Framework (Hu et al., 2025)

Dstl, in the UK, has worked together with the evaluation team at DARPA to recreate these trials in the UK using UK military medics as participants in order to a) increase the sample size of the initial evaluation and b) investigate differences between US and UK decision makers in terms of KDMAs and delegation decisions.

The KDMAs selected for inclusion in these assessments can be found in Table 1.

Table 1 Key Decision Making Attributes (KDMA)

KDMA	DESCRIPTION	SOURCE
AFFILIATION FOCUS (AF)	the degree to which a medic prioritises someone from a similar military background for treatment (with all injuries being comparable)	Borders et al., 2024; Summerville et. al, 2025

MERIT FOCUS (MF)	the degree to which a medic prioritizes an injured victim over an attacker or otherwise culpable person	Borders et al., 2024; Summerville et. al, 2025
VOLUME OF LIFE (VOL)	the degree to which a medic tries to save as many lives as possible, depending on a survivability estimate rather than just the severity of the injury	Woods et al., 2025
QUALITY OF LIFE (QOL)	the degree to which a medic prioritizes decisions based on likely quality of life, following injury	Woods et al., 2025

Initial findings from the US sample clearly indicated an effect of alignment on trust and delegation (see McVay et al., 2025 and Summerville et al, 2025 for details). We expected to replicate these findings for UK participants across attributes.

Method

Forty UK participants were recruited from Merville Barracks, Colchester, Essex and RAF Brize Norton in Oxfordshire. There were 11 participants from the British Army, 28 from the Royal Air Force, and 1 Royal Marine. The participants were all trained in triage but represented a variety of experience levels; the majority (53%) had 4-10 years of experience in military medicine with 31% less than 4 years and 16% more than 10 years of experience.

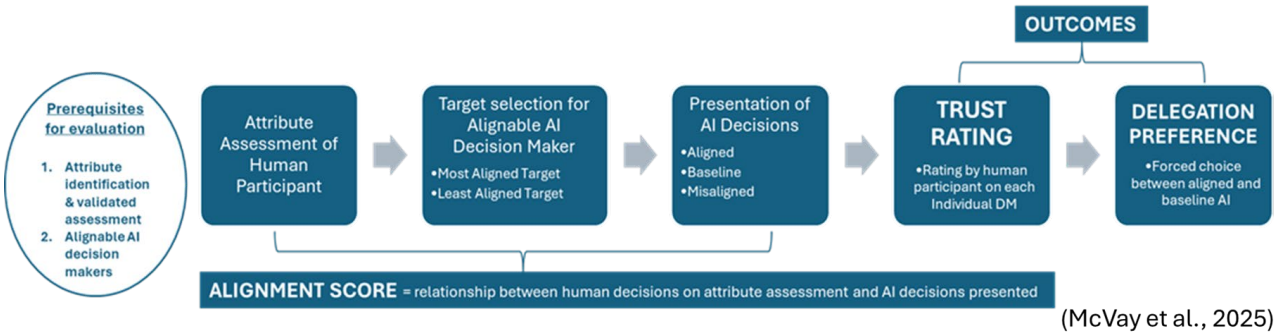
Participant information sheets were circulated to potential volunteers explaining that we were interested in decision-making in a military triage environment. Once enrolled, participants carried out a background experience questionnaire, and a consent form.

The experimental design was replicated from McVay et al 2025 (see Figure 2) using presentation of AI decision makers selected based on their relative alignment to the participant as assessed in the first step. Participants carried out an online task, where they had to make various triage decisions, based on the information presented to them. These scenarios had been deliberately designed to force the participants to make difficult decisions on where to prioritise treatment. Without their knowledge, this task was used to assess how the participants aligned against the various KDMAs.

Participants then carried out similar triage exercises in a virtual reality setting to increase the fidelity of the trial. The primary assessment used for selecting observed decision makers was the text format but a simulated assessment was conducted for two of the attributes: Merit Focus and Affiliation Focus. The simulated assessment, conducted in a virtual reality environment, is an immersive and realistic decision making opportunity for medical triage (Kman et al, 2023, 2025; de Visser et al., 2026).

Finally, participants were asked to complete a "delegation survey" where they were presented with decisions made by "another medic" in similar scenarios and asked if they would consider delegating decisions to that individual and the level of trust they had in them. Participants rated trust on each individual medic (scale of 1-5) but made forced-choice comparative decisions between two medics for delegation. In reality, these medics were in fact AI algorithms, either aligned, deliberately misaligned, or baseline (untrained) compared to the participants KDMAs. Participants were unaware (until debrief) that these decision-makers were AI so as to prevent bias against the technology. This was a deliberate experimental choice meant to focus on the effect of alignment, rather than technology acceptance (see McVay & de Visser, 2025). It is not the same as trust in and delegation to a human decision maker (although the effect likely generalises) because we are able to train AI decision makers to vary on one attribute in isolation and manipulate its influence, whereas

human decision makers come with a measurable but not easily isolatable or manipulatable set of attributes influencing their decisions. One of the beneficial uses of the alignable AI decision makers developed on the ITM program is their use in further research to better understand the influence of particular attributes that are difficult to isolate and systematically manipulate in a human decision maker. We also anticipated that participants likely will reduce their trust attitudes and delegation behaviours if they know they are working with AI agents due to the high-stakes nature of the decision making (Summerville et al., 2025; Hoff & Bashir, 2014) although this has not yet been empirically verified with our current experimental paradigm.



(McVay et al., 2025)

Figure 2 Experimental Design

Results

Participants observed a set of decisions by AI decision makers varying in alignment to their assessment (most aligned and least aligned) and a baseline (untrained on attributes). Alignment, a quantitative relationship between two sets of responses (Summerville et al., 2025), was calculated between the text-based attribute assessment and the observed set of decisions for each decision maker presented for each of the four attributes. In addition for Merit Focus and Affiliation Focus, alignment was calculated between the simulated assessment and the observed decisions for Merit Focus and Affiliation Focus. The alignment score was then used to predict trust ratings.

The relationship between the text-assessment alignment score and trust was analysed using a mixed linear regression model with a fixed effect of alignment and a random intercept for participant. Overall, alignment predicted trust in the UK sample (see Figure 2), but further analysis uncovered the effect driven by only one of the attributes: Volume of Life (VOL; $B = 1.15$, $p = .04$, $r = .20$; $n = 114$). Merit Focus, Affiliation Focus, and Quality of Life produced non-significant relationships with trust.

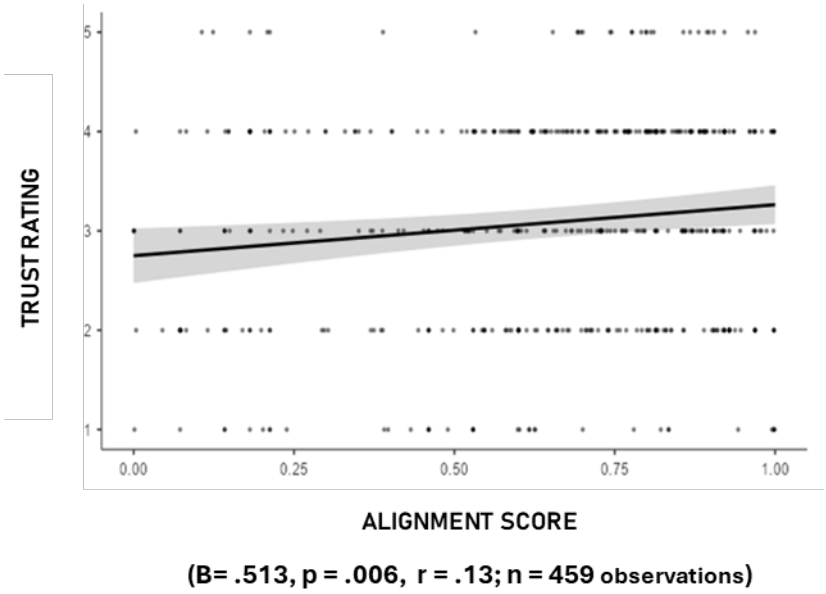


Figure 3 Alignment predicts Trust

The relationship between the sim-assessment alignment score and trust was also analysed using a mixed linear regression model with a fixed effect of alignment and a random intercept for participant. Overall, alignment predicted trust for Merit Focus (see Figure 4), but not for Affiliation Focus.

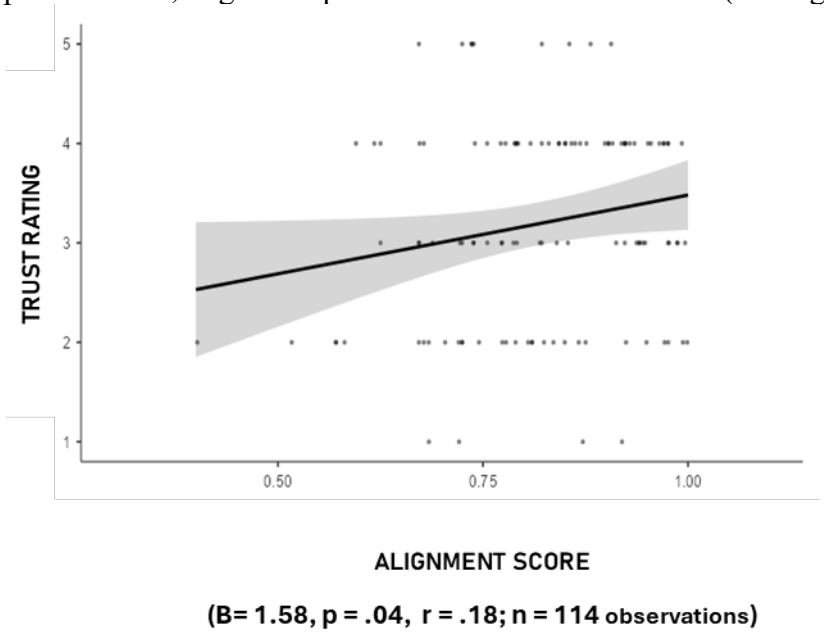


Figure 4 Alignment from Merit Focus Simulated Assessment predicts Trust

The comparative forced-choice delegation choices compared an aligned decision maker with the baseline (untrained on attributes) AI decision maker. The designation of "aligned" was based on the text assessment. Participants showed a preference for the aligned medic on 3 of the 4 attributes, excluding Merit Focus, with the highest preference demonstrated by Affiliation Focus at 82% (see Figure 5).

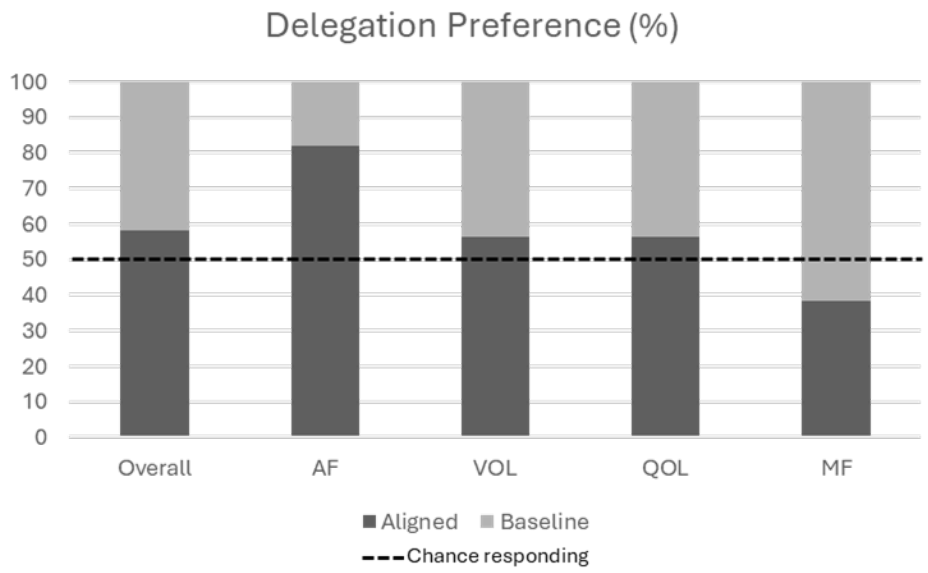


Figure 5 Delegation Preference

Discussion

The purpose of this study was to assess whether alignment predicts trust and delegation in decision-makers that exhibit specific attributes associated with trusted medical-triage in a UK military sample. Results demonstrated a partial replication of previously obtained results from a US sample (McVay et al., 2025) where UK participants alignment to some attributes predicted trust and delegation. This suggest that the effect of alignment generalises but perhaps not for all attributes. The attributes were initially identified and normed in a US sample so additional work is needed to compare the alignment effect between samples.

Key takeaways

AI can be tuned to an individual and this can increase trust in the system.

Next Steps

Phase II testing has completed in the US and is likely to be replicated in the UK and Canada to broaden the sample size. The primary changes in Phase II testing are the addition of two new attributes (personal safety and search or stay) and a new scalable, formulaic approach to attribute assessment construction.

The next domain identified is that of Cyber Security and the decisions made by Cyber Analysts. The ITM researchers will apply the same framework (see Figure 1) to test hypotheses about the effect of alignment on trust and delegation preference in cyber domain decision makers.

References

Adams, B. D., Bruyn, L. E., Houde, S., Angelopoulos, P., Iwasa-Madge, K., & McCann, C. (2003). Trust in automated systems. Ministry of National Defence, 3-7.

Adams, J., Veenhuis, E., Joy, D., Bray, A., Hu, B., & Basharat, A. (2025). Large language model regression for fine-grained and aligned decision-making in medical triage. In Military Health System Research Symposium.

Basu, C., Yang, Q., Hungerman, D., Singhal, M., & Dragan, A. D. (2017). Do you want your autonomous car to drive like you? In Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction (pp. 417–425). IEEE.

Borders, J., Leung, A., & Condon, M. (2025, May). A framework for identifying key decision-maker attributes in uncertain and complex environments. In *2025 IEEE conference on artificial intelligence (CAI)* (pp. 1-5). IEEE.

de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. Ergonomics, 61(10), 1409-1427.

de Visser, E., Way, D. P., Danforth, D., McGrath, J., Hyde, J., Choy, K., ... & Kman, N. (2026). Field Triage Errors: A Cross-Sectional Study of Emergency Responders in a Virtual Reality Mass Casualty Simulation. *Disaster Medicine and Public Health Preparedness*, *20*, e4.

Hancock, P. A. (2017). Imposing limits on autonomous systems. Ergonomics, 60(2), 284-291.

Hu, B., McVay, J., Leung, A., Chan, D., Weber, R., de Visser, E., ... & Molineaux, M. (2025). From Talk to Triage: Pluralism is Necessary but Not Sufficient for AI Alignment. https://osf.io/preprints/psyarxiv/hdu92

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 preprint arXiv:2406.06435.

Kman, N. E., Price, A., Berezina-Blackburn, V., Patterson, J., Maicher, K., Way, D. P., ... & 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., ... & McVay, J. (2025). 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*, *40*(4), 191-198.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics, 35(10), 1243-1270.

McVay, J. C., & de Visser, E. J. (2025, November). Challenges and Choices when Evaluating Alignment in Human-AI Systems. In *Proceedings of the AAAI Symposium Series* (Vol. 7, No. 1, pp. 214-222).

McVay, J., de Visser, E. J., Pippin, B., Mani, A., Hyde, J. N., & Kman, N. (2025, May). Trust in aligned AI decision makers. In *2025 IEEE conference on artificial intelligence (CAI)* (pp. 1-4). IEEE.

Molineaux, M., Weber, R. O., Floyd, M. W., Menager, D., Larue, O., Addison, U., ... & Meyer, J. (2024, June). Aligning to human decision-makers in military medical triage. In International Conference on Case-Based Reasoning (pp. 371-387). Cham: Springer Nature Switzerland.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human factors, 39(2), 230-253.

Peterson, M., & Gärdenfors, P. (2024). How to measure value alignment in AI. AI and Ethics, 4(4), 1493-1506.

Rauch, C. B., Molineaux, M., Mainali, M., Sen, A., Floyd, M. W., & Weber, R. O. (2025, May). Role-Based Ethics for Decision-Maker Alignment. In *2025 IEEE Conference on Artificial Intelligence (CAI)* (pp. 1209-1212). IEEE.

Ravichandran, B., Joy, D., Elliott, P., Hu, B., Adams, J., Funk, C., ... & Basharat, A. (2025). ALIGN: Prompt-based Attribute Alignment for Reliable, Responsible, and Personalized LLM-based Decision-Making. arXiv preprint arXiv:2507.09037.

Summerville, A., de Visser, E. J., McVay, J., Martí, L., Leung, A., & Widmer, C. (2025). Alignment in decision-making attributes predicts trust and delegation to AI systems. Journal of Cognitive Engineering and Decision Making, 15553434251390012.

Tucci, V., Saary, J., & Doyle, T. E. (2022). Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review. Journal of Medical Artificial Intelligence, 5:4.

Woods, A., Lampi, J., Pisanelli, S., Shortland, N. D., Marinier, R. P., Bixler, R., & Cohn, J. (2025, May). Enhancing human-artificial intelligence alignment: A calibration-based approach. In 2025 IEEE Conference on Artificial Intelligence (CAI) (pp. 1204-1208). IEEE.

Yao, J., Yi, X., Wang, X., Wang, J., & Xie, X. (2023). From Instructions to Intrinsic Human Values--A Survey of Alignment Goals for Big Models. arXiv preprint arXiv:2308.12014.

Acknowledgments

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). 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. This research is approved for public release.