

Stanford

Geopolitics, Technology,
and Governance
Center for International Security and Cooperation



THE FUTURE OF DECISION_MAKING CONFERENCE
28th October 2026

AGENDA

Welcome and Framing	Brad Boyd, Zena Wood	09:00
Keynote Address	Lt. Gen. Jack Shanahan	09:15
Research Presentation	Paul Larcey	10:00
Break		10:35
Panel: AI and the Path to Cognitive Surrender	Brad Boyd, Lt. Gen. Jack Shanahan, Nand Mulchandani	10:50
Military Decision-making and AI	LTC Indrek Sarap	11:40
Lunch		12:15
Command by Algorithm? AI-DSS and the Allocation of Battlefield Risk	Andrew Bell, Jacquelyn Schneider	13:00
When Uncertainty Expands Authority: How AI Workflows Can Reduce Scrutiny	Eli Talbert	13:35
Preventing Free Rein and Nominal Reign: Disaggregated Civilian Harm Reporting	Daniel Lucuix	14:10
Break		14:45
AI Supported Professional Judgment: Explaining Situated Human-Machine Decision-making	Tim Jones	15:00
The Aid Sector at Machine Speed: AI and the Future of Humanitarian Decision-making	Rachel George	15:35
Self-determined nudging and boosting cognitive competences to improve decision-making in the security-related sector: Exploratory sketch towards a CoBoNu framework	Anja Dannewitz	16:10
Closing Remarks	Brad Boyd, Zena Wood	16:45

Welcome

Welcome to the Future of Decision-Making Project's conference.

As AI-enabled systems become increasingly integrated into high-stakes decisions - from military operations and intelligence analysis to crisis management and clinical care - we face important questions about how these technologies shape human judgment, accountability, and responsibility.

Today's conference brings together scholars, practitioners, and operational experts to explore these questions, assess the current state of the field, and identify priorities for future research.

On behalf of the Future of Decision-Making Project at Stanford's Center for International Security and Cooperation, thank you for joining us. We look forward to a thoughtful and engaging discussion.

Brad Boyd
Founder and Senior Advisor, FODM

Keynote: Advanced AI in Military Operations: Epistemic Dependence or Epistemic Capture? What's the Dif- ference and When Does One Become the Other?

Military decision-makers have always relied on systems and staffs that know more than they do. Advanced AI raises a harder question: not just who decides, but who frames the decision. As Desireé Lorell puts it, “The machine didn’t decide, but it decided what was decidable.” Yet commanders still answer for the outcome. Shanahan draws a sharp line between dependence and capture. Dependence is when we cannot recreate what the machine knows. Capture is when we can no longer challenge how it has framed what we know. He examines where that line falls, whether advanced AI fundamentally changes the answer, and what an accountable chain of human assurance must look like as AI moves into the center of military decision-making. The goal is not to keep the machine out of the decision, but to keep the human capable of contesting it.

Lt. Gen. Jack Shanahan



Panel: AI and the Path to Cognitive Surrender



Lieutenant General John (Jack) N.T. Shanahan, United States Air Force, Retired, retired in 2020 after a 36-year military career. In his final assignment he served as the inaugural Director of the U.S. Department of Defense (DoD) Joint Artificial Intelligence Center (JAIC). Jack served in a variety of operational and staff positions in various fields including flying, intelligence, policy, and command and control. He commanded at the squadron, group, wing, Agency, and Numbered Air Force levels. As the first Director of the Algorithmic Warfare Cross-Functional Team (Project Maven), Jack established and led DoD's pathfinder AI fielding program charged with bringing AI capabilities to intelligence collection and analysis.



Nand Mulchandani is a serial entrepreneur with a passion for public service and a love for small, fast-moving teams focused on disrupting big markets. He was most recently the chief technology officer (CTO) at the Central Intelligence Agency and the CTO and acting director of the US Department of Defense's Joint Artificial Intelligence Center. Prior to pivoting to government service, Mulchandani was cofounder and CEO of multiple successful start-ups—Oblivix Inc. (acquired by Oracle), Determina (acquired by VMware), OpenDNS (acquired by Cisco), and ScaleXtreme (acquired by Citrix). He was a Senior Executive at VMware and Citrix and started his career as a Compiler Engineer at Sun Microsystems. Mulchandani has a degree in Computer Science and Math from Cornell University, a Master of Science in Management from the Stanford Graduate School of Business, and a Master of Public Administration from Harvard's John F. Kennedy School of Government.



Brad Boyd is the Founder and Senior Advisor of FODM. Boyd's research interests focus on the integration of emerging technology into warfare and national security. Boyd is particularly interested in the rise of automation and autonomy in military systems, to include decision making, planning, information operations, enterprise operations, and weaponry. Boyd also looks at the way US, Chinese, and Russian military integration of emerging technology affects military, economic, and social stability. Prior to joining Hoover, Boyd served as a defense and foreign policy advisor to Senator Angus King and as Senator King's representative to the Cyberspace Solarium Commission. Boyd also served as the director of AI-enabled warfighting capability development at the US Department of Defense's Joint Artificial Intelligence Center and as the director of General Mark A. Milley's coordination group.



RESEARCH PRESENTATIONS

Command by Algorithm? AI-DSS and the Allocation of Battlefield Risk

Andrew Bell, Ph.D., and Jacquelyn Schneider, Ph.D.
Corresponding author: ambell@stanford.edu

Introduction

On July 3, 1988, operating in the hazardous waters of the Persian Gulf, the USS Vincennes (CG-49) detected a potentially hostile air contact that appeared to be heading toward the ship. In the high-stakes, time-compressed threat environment, the ship's automated Aegis Combat System tracked the contact, issuing machine-generated cues that, through crew misinterpretation, led to the narrowing of the range of tactical alternatives considered by the crew. Under time pressure, and with the contact subsequently classified as hostile, the Vincennes' commander gave the order to launch (Rochlin, 1990). The resulting shoot-down of Iran Air Flight 655, a commercial airliner, killed 290 civilians, inflaming the already tense geopolitical relations between the U.S. and Iran. The incident also illustrated how automated systems can structure perception and decision-making in ways that lead to increased risk of command error and civilian harm under operational pressure.

Decades later, automation has made a revolutionary leap on the modern battlefield; AI-enabled weapons, communications, and decision-making tools now proliferate across conflict. But while AI is often framed as a transformative technology that can revolutionize the speed, scale, and lethality of modern warfare, its more consequential effects may lie in its implications for military command. AI is not only changing how militaries fight; it is reshaping how decisions about the use of violence are generated and executed.

AI and the Combatant's Trilemma

AI-human command decisions aren't just

about making the "best" decisions, they represent a trade-off between military advantage, the protection of friendly forces, and the harm imposed on civilians in war. The tension between the values of this "combatant's trilemma" often forcing commanders to choose the "lesser evil" in the conduct of military operations (Bell, 2022). In military operations, the balancing of these three competing values depends not only on commanders' preferences but also on command relationships—the processes and dynamics through which commanders structure decision-making, communicate intent, and direct and constrain the use of force.

The introduction of AI into the combatant's trilemma poses obvious risks. Research on LLMs within scenarios and games finds that agents are more likely to use violence and earlier in a conflict than humans (Lamparth et al., 2024). Similarly, a study on LLMs and the laws of armed conflict concluded that LLMs consistently violated legal principles on the use of force (Drinkall 2025). This research suggests that, when balancing the trilemma, existing models may overemphasize military advantage or protecting friendly forces over civilian harm—thus increasing the chance that AI-enabled warfare makes war more dangerous for civilians.

But there is significant variation in agent behavior across models and user prompts (Payne, 2026) and militaries across the world insist that humans will retain final decision-making. How then might AI affect the combatant's trilemma—regardless of model tendencies in a world where militaries are not directly substituting AI recommendations for human decision-making? We argue that AI-enabled decision support systems

(AI-DSS) significantly reshape commanders' decisions about the trilemma by: 1) reorganizing how judgment is produced within military organizations; 2) compressing time and decision-making processes; and 3) reshaping how command knowledge is converted into judgment as part of operational art.

How military organizations build these command processes may ultimately determine whether AI makes war more or less dangerous for civilians.

1. Organization: AI Reshapes Staff Cognition and Command Processes

Warfighting is increasingly a hybrid form of human-machine judgment rather than a purely human or machine process (Schneider & Macdonald, 2025; Jensen et al., 2023). As Boyd and Saade note, AI-enabled warfighting is a hybrid cognitive process in which machines not only execute human decisions but also structure the information, options, and judgments from which those decisions emerge (Boyd and Saade 2025).

As AI systems generate courses of action, prioritize targets, and structure operational choices, they increasingly shape the cognitive environment in which staffs operate (Horowitz et al. 2024; Chapa and Cantu 2025). Emerging evidence from military experimentation suggests that these systems can anchor thinking within command environments. In planning cycles, staff personnel may treat AI-generated options as an initial baseline, adjusting around them rather than developing alternatives independently (Wilcox & Metcalf, 2026). A broader body of research shows that automation bias and anchoring effects can lead individuals to rely on machine-generated outputs, even when those outputs are flawed. If AI systems increasingly define the initial set of options, they shape the decision space before commanders engage with it (Denzel, 2026).

2. Time: Compressing Space for Human Judgment

A central promise of military AI is speed (Iraqi, 2024). AI in military decision-making drastically speeds up the time in which operational plans can be developed, target packages can be generated, and targeting cycles can be executed. AI in operational planning can compress decision cycles from hours or days to seconds (Jensen et al., 2025).

While these systems increase the tempo of decision-making, they can also reduce the cognitive space and opportunity for independent human judgment. As the space for command reflection narrows, decision-making can become more reactive than generative—less about developing and weighing options than selecting among those generated by an opaque software process.

Israel's use of the Lavender system in Gaza illustrates the point: according to investigative reporting, the AI-DSS target generation system reportedly led Israeli officers to spend less than half a minute on vetting targets, leading these officers to characterize their role as a "stamp of approval" in a targeting process largely driven by AI. Such reports suggest that in such high-tempo operations, while formal human authorization remained in place, the time and space for independent human judgment narrowed sharply (Iraqi, 2024).

Under these conditions, the values of mission accomplishment and force protection—objectives that lend themselves to rapid assessment and salient mission effects—are more likely to dominate. Civilian protection, which often requires additional deliberation and restraint, risks being de-emphasized. Speed, in other words, reshapes the conditions under which tradeoffs are made—and the direction in which the combatant's trilemma tensions are resolved.

3. Knowledge: Opaque AI and the Externalization of Judgment

A third shift is in many ways less visible but potentially the most consequential. Advanced AI systems can generate recommendations that are difficult to fully explain. Commanders may be asked to act on outputs whose internal reasoning they cannot independently reconstruct—what some have described as reliance on “alien oracles” (Wehmeyer, 2025).

Opaque AI systems complicate this process in three ways. First, they weaken contestability: If commanders cannot understand how a system arrives at an operational recommendation, it becomes harder to interrogate the assumptions embedded in that recommendation—assumptions about target value, civilian presence, or acceptable risk. More importantly, an AI system may not just provide an answer; it can shape or even determine what elements are counted as relevant in the decision-making process.

Second, opacity can generate false confidence (Macdonald & Schneider, 2018). Machine outputs often appear precise and comprehensive, encouraging deference, especially under time pressure. In particular, researchers find that AI imbues overconfidence, especially in tasks the individual is otherwise not expert in (Fernandes et al 2025). Such overconfidence can further create potential for error in decision-making.

Third, and most important, opaque systems lead to the externalization of judgment. With AI-DSS, commanders may retain formal authority (Horowitz, 2025) but the substantive reasoning behind decisions is likely embedded, at least partially, in the system itself (Chapa, 2024). The mere presence of a human decision-maker does not resolve this problem; even human-in-the-loop systems can distort judgment when machine outputs shape decisions on grounds that cannot be adequately scrutinized (Chapa, 2024). This dynamic also contributes to what can

be characterized as “distributed moral distance” (Chapa, 2022). Responsibility for the use of force is spread across a network of actors—engineers, analysts, staff officers, and machines—such that, with AI-assisted decision-making, no single individual fully experiences themselves as the decisive moral agent.

Within the trilemma, this matters because the critical question is not only how much harm is predicted, but how competing values are weighed. AI may improve what militaries can know, but it can also reshape how that knowledge is converted into judgment and who, in practice, is exercising it (Jensen & Tadross, 2023).

Conclusion

AI-DSS have significant implications for commanders’ role in warfare and the nature of violence inflicted on the battlefield. Ultimately, we do not claim that AI-DSS necessarily or inherently increase civilian harm. Rather, we argue that the integration of AI-DSS can reshape how commanders weigh the competing values of the combatant’s trilemma, leading to prioritization of more quantifiable, machine-legible mission and threat values over risks to civilians. The central challenge, therefore, is not merely to enable faster, more lethal military operations, nor to keep commanders in the loop during these operations, but to ensure that civilian populations remain fully visible within the judgments AI-DSS enable commanders to make on the battlefield.

References

- Bell, A. (2022). Combatant Socialization and Norms of Restraint: Examining Officer Training at the US Military Academy and Army ROTC. *Journal of Peace Research*, 59(2), 180–196. <https://doi.org/10.1177/00223433211010861>
- Chapa, J. (2024). Military AI Ethics. *Journal of Military Ethics*, 23(3–4), 306–321. <https://doi.org>

/10.1080/15027570.2024.2439654

Chapa, J. (2022). Is Remote Warfare Moral?: Weighing Issues of Life and Death from 7,000 Miles. PublicAffairs.

Chapa, J. O., & Cantu, S. (2025, October 14). Human-Machine Planning: AI Lessons from the Marine Command and General Staff College. War on the Rocks. <https://warontherocks.com/human-machine-planning-ai-lessons-from-the-marine-command-and-general-staff-college/>

Denzel, C. (2026, February 26). AI Is Being Misunderstood as a Breakthrough in Planning. It's Not. War on the Rocks. <https://warontherocks.com/cogs-of-war/ai-is-being-misunderstood-as-a-breakthrough-in-planning-its-not/>

Drinkall, T. (2025). Red lines and grey zones in the fog of war: Benchmarking legal risk, moral harm, and regional bias in large language model military decision-making (arXiv Preprint arXiv:2510.03514). arXiv. <https://doi.org/10.48550/arXiv.2510.03514>

Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., ... & Welsch, R. (2025). AI makes you smarter but none the wiser: The disconnect between performance and metacognition. Computers in Human Behavior, 108779.

Horowitz, M. (2025, May 22). Autonomous Weapon Systems: No Human-in-the-Loop Required, and Other Myths Dispelled. War on the Rocks. <https://warontherocks.com/autonomous-weapon-systems-no-human-in-the-loop-required-and-other-myths-dispelled/>

Horowitz, M., Kahn, L., Macdonald, J., & Schneider, J. (2024). Adopting AI: How Familiarity Breeds Both Trust and Contempt. AI & Society, 39(4), 1721–1735. <https://doi.org/10.1007/s00146-023-01666-5>

Iraqi, A. (2024, April 3). 'Lavender': The AI Machine Directing Israel's Bombing Spree in Gaza. +972 Magazine. <https://www.972mag.com/lavender-ai-israeli-army-gaza/>

Jensen, B., & Kwon, J. S. (2025, March 10). The U.S. Army, Artificial Intelligence, and Mission Command. War on the Rocks. <https://warontherocks.com/the-u-s-army-artificial-intelligence-and-mission-command/>

Jensen, B., & Tadross, D. (2023, April 12). How Large-Language Models Can Revolutionize Military Planning. War on the Rocks. <https://warontherocks.com/how-large-language-models-can-revolutionize-military-planning/>

Jensen, B., Tadross, D., & Strohmeyer, M. (2025, April 23). Agentic Warfare Is Here. Will America Be the First Mover? War on the Rocks. <https://warontherocks.com/agentic-warfare-is-here-will-america-be-the-first-mover/>

Jensen, B., Whyte, C., & Cuomo, S. (2023, July 20). The Future of Algorithmic Warfare: Fragmented Development. War on the Rocks. <https://warontherocks.com/the-future-of-algorithmic-warfare-fragmented-development/>

Lamparth, M., Corso, A., Ganz, J., Mastro, O. S., Schneider, J., & Trinkunas, H. (2024). Human vs. machine: Behavioral differences between expert humans and language models in wargame simulations. In Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (Vol. 7, No. 1, pp. 807–817). <https://doi.org/10.1145/3630106.3658971>

Macdonald, J., & Schneider, J. (2018, February 5). Trust, Confidence, and the Future of Warfare. War on the Rocks. <https://warontherocks.com/trust-confidence-future-warfare/>

Payne, K. (2026). AI arms and influence: Frontier models exhibit sophisticated reasoning in simulated nuclear crises (arXiv Preprint arXiv:2602.14740). arXiv. <https://doi.org/10.48550/arXiv.2602.14740>

org/10.48550/arXiv.2602.14740

Rochlin, G. I. (1990). Iran Air Flight 655 and the USS Vincennes. University of California, Berkeley.

Schneider, J., & Macdonald, J. (2025). The Hand Behind Unmanned: Origins of the US Autonomous Military Arsenal. Oxford University Press.

Wehmeyer, J. (2025, September 26). Alien Oracles: Military Decision-Making with Unexplainable AI. War on the Rocks. <https://warontherocks.com/alien-oracles-military-decision-making-with-unexplainable-ai/>

Wilcox, A. B., & Metcalf, C. (2026). AI Command and Staff—Operational Evidence and Insights from Wargaming. Military Strategy Magazine, 10(4). <https://doi.org/10.64148/msm.v10i4.1>

When Uncertainty Expands Authority: How AI Workflows Can Reduce Scrutiny

Eli Talbert, PhD

The MITRE Corporation

Corresponding author: etalbert102@gmail.comz

When a Defensible Label Removes a Safeguard

At 2:17 a.m., a pale shape moves along a secondary road north of friendly positions. Atmospheric distortion smears its outline. On the operations center display, it occupies just twelve by eighteen pixels. A tracking box closes around it, and a label appears: VEHICLE. The track remains in the routine queue. No higher-resolution sensor is retasked, and no reviewer is alerted. Minutes later, the object passes beyond view.

The operations team never sees the model's underlying estimate: "pickup-like silhouette, 42% confidence," with alternatives including a civilian pickup, a modified truck, and a technical—an improvised armed truck—carrying a crew-served weapon. To avoid false specificity and liability risk, the vendor reports "vehicle" below 60 percent while retaining alternatives internally. A legacy interface then transmits `object_class = vehicle` but cannot carry confidence or alternatives. The generic label keeps the track in routine processing. An unresolved specific report would trigger observation, sensor retasking, or human review.

The object is a technical transporting a crew-served weapon. Its presence was detected; its threat was not. It proceeds without the added scrutiny that could have provided warning. Without knowing the confidence threshold, the opposing force watches which movements draw additional sensing or human attention. By varying distance and exposure under different lighting, it can learn when ambiguity yields "vehicle" and routine processing. The attacker has

not caused a false prediction. It has induced the system to generalize—and proceed.

The Decision Space as an Attack Surface

This essay advances a testable hypothesis: holding model output constant, a workflow that hides a policy-relevant alternative and routes on a broad fallback label gives ambiguous cases less scrutiny—and more permission to proceed—than one that preserves it. Uncertainty becomes dangerous when confidence rules, lossy interfaces, and permissive routing remove safeguards.

An AI system's decision space is the actions people and machines present, recommend, allow, or take. A distinction matters when different answers call for different evidence, safeguards, or permission. In this specific case, no one team designs the whole chain. Instead, evidence passes through choices made by different teams:

Evidence → model estimate → confidence rule → interface representation → routing rule → available or automatic actions

Model developers set confidence rules. Interface designers choose fields to send. Operations leaders decide which categories trigger intervention. Together, these choices create functional and workflow authority. The end user has operational authority only if the track shows the distinction and lets them stop, reverse, or challenge the outcome.

The vendor's "vehicle" fallback avoids false specificity while retaining the estimate. The legacy interface drops both confidence and

alternative-label fields. Operations treats the category as a reason not to escalate, not just as a description of the evidence. The key question is whether this broader but accurate label should clear the track from scrutiny.

Unlike targeted misclassification, where altered inputs produce wrong classifications (Eykholt et al., 2018), the attacker seeks a broad fallback that continues routine processing and avoids escalation. Selective prediction lets a model withhold an answer when uncertain (Franc & Průša, 2019), but it helps only if the workflow adds scrutiny or limits high-stakes action. A correct broad category and a well-calibrated confidence score still do not decide what should happen next.

If an opposing force exploits the link between a defensible output and a permissive workflow, the flaw becomes adversarial decision-space manipulation. The attack succeeds because the track leaves scrutiny before anyone can collect more information or intervene. Better sensors and calibration may improve the estimate, but they cannot tell us whether the workflow grants the right permissions.

How Ambiguity Shifts Authority

The design flaw becomes an attack only if an opposing force can create ambiguity, observe a repeatable pattern in the response, trigger it again, and benefit.

Military history shows this kind of adaptation. The Taliban observed Coalition forces, shared discovered vulnerabilities, and deduced limits in national rules of engagement (Afsar & Samples, 2008; Ryan, 2007). These examples do not establish that an AI fallback has been exploited. They show that opposing forces can learn rules from repeated responses without seeing the rules themselves. The design becomes exploitable only when a repeatable link connects broader labels to less scrutiny.

Human teams can adapt when patterns become visible. Here, however, automation de-

termines what receives human attention. When it routes ambiguous cases away from review and records a defensible broad category, people may never see the pattern.

Component tests can miss this shift in authority. If an evaluation rewards broad categories or penalizes overly specific errors, accuracy can rise even while armed vehicles trigger escalation less often. Confidence may remain calibrated, and every component may follow its rule. Each team may meet its metric while the combined system declines to escalate on weaker evidence.

The vulnerability spans components and organizational boundaries. The vendor reports a broader category to avoid overclaiming and liability. The integration team preserves compatibility with legacy software. Operations limits false alarms by keeping broad categories in the routine queue. Each choice can meet its owner's metric, yet no one owns the combination. Together, they turn a cautious fallback into permission to proceed.

That permission can bypass review when routing rules keep the case from a reviewer. A reviewer has operational authority only if they can see the suppressed alternatives and retask, stop, or reroute the workflow. A review right that never activates is not control.

A Governance Rule for Uncertainty

The design rule is simple:

When evidence for a distinction that affects policy becomes weaker, that alone should not reduce scrutiny or allow more action without review.

In the road scenario, uncertainty should prompt longer observation or a sensor retask, not reduce scrutiny. Elsewhere it might justify isolating a network segment. If delay poses greater danger, an approved emergency rule may allow high-stakes action with an audit trail. Its justification is the

danger of delay, not a broader label. High-stakes action must not expand by accident or without fresh approval. Severity, reversibility, and automation govern the response.

Because the rules create the vulnerability together, evaluators must test the whole workflow, not just its parts. The first phase adds an audit of routing rules to tests of accuracy, calibration, and robustness. Evaluators run one uncertain output through two policies: one preserves the unresolved alternatives, while the other falls back to the broader category. With the model output fixed, any difference comes from the policy. The comparison shows whether the fallback changes scrutiny, permissions, automation, or the reviewer's ability to intervene.

The second phase is an adversarial stress test. Evaluators add noise or partly block the image, then repeat the test. An independent red team watches the system's responses to find and repeatedly trigger the fallback. If weaker evidence lets a case proceed with less scrutiny, the system fails unless an approved rule ties that change to another condition, such as the danger of delay.

Automatic human review is not enough if an attacker can flood the queue or repeatedly cross its thresholds. The audit should test the review mechanism under load. Triage should favor reversible ways to gather information and reserve review for cases where judgment can change what the system may do.

Each record should link the reported category to hidden alternatives and the rule it triggered. It should also record later interventions or overrides and identify the system version. A named owner must be able to change the thresholds, displays, and routing rules together. Without that system-wide power, the organization can see the problem but cannot control it.

The audit asks a practical question: can accountable people see and change how uncertainty affects what the system may

do? It turns calls for context-specific oversight and distinct human roles (Tabassi, 2023), plus warnings about AI outrunning "human firewalls" (U.S. Air Force, 2025), into a test of who actually has authority.

AI-generated prose can create the same problem in civilian workflows. A language model compresses conflicting evidence and gaps into one summary. The interface hides the alternatives, and a later rule treats the summary as settled. People may trust model explanations too much, and longer explanations can raise confidence without improving accuracy (Steyvers et al., 2025). The audit asks whether this sequence reduces scrutiny or allows more to proceed.

Limits

The road case is hypothetical, not evidence of an exploit in a real system. It follows a single model output through its confidence rule, interface, and routing. Historical cases show how opponents learn from repeated operational responses. That distinction matters because the same weak design can exist before an opponent discovers it, while its operational risk depends on whether exploitation is practical. The vulnerability becomes an attack when someone can create, observe, repeat, and benefit from the combined effect.

Meaningful human control depends on how the parts work together: when the confidence threshold, interface, and routing rule combine, does weaker evidence relax safeguards before a person acts? The system need not miss the object to miss the reason to hesitate.

References

Afsar, S. A., & Samples, C. A. (2008). The evolution of the Taliban [Master's thesis, Naval Postgraduate School]. Calhoun: The NPS Institutional Archive. <http://hdl.handle.net/10945/4101>

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1625–1634). IEEE Computer Society. <https://doi.org/10.1109/CVPR.2018.00175>

Franc, V., & Průša, D. (2019). On discriminative learning of prediction uncertainty. In K. Chaudhuri & R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning (Vol. 97, pp. 1963–1971). PMLR. <https://proceedings.mlr.press/v97/franc19a.html>

Ryan, M. (2007). The other side of the COIN: Reconstruction operations in southern Afghanistan. *Australian Army Journal*, 4(2), 125–143. <https://researchcentre.army.gov.au/library/australian-army-journal-aaj/volume-4-number-2/other-side-coin-reconstruction-operations-southern-afghanistan>

Steyvers, M., Tejeda, H., Kumar, A., Belém, C. G., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2), 221–231. <https://doi.org/10.1038/s42256-024-00976-7>

Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>

U.S. Air Force. (2025, April 8). Air Force doctrine note 25-1: Artificial intelligence (AI). https://www.doctrine.af.mil/Portals/61/documents/AFDN_25-1/AFDN%2025-1%20Artificial%20Intelligence.pdf

Preventing Free Rein and Nominal Reign: A Disaggregated Civilian Harm Reporting Framework for AI-Assisted Targeting

Daniel Lucuix

Stanford University

Corresponding author: dvp@stanford.edu

I. Introduction

Conversations around military decision-makers' use of artificial intelligence (AI) have exploded since the U.S. Department of Defense began integrating AI Decision Support Systems (AI-DSS) into its military operations, particularly in Iran. Initial reporting would suggest that the U.S. has extensively utilized Palantir's Maven Smart System to identify targets for Operation Epic Fury, demonstrated by its peak daily usage rising 4,425 percent (Freedberg, 2026). AI-DSS are computerized tools that generate analyses, recommendations, and predictions to inform human decision-making, using machine learning to work over data sets too large and complex for traditional systems (Holland Michel, 2024, p. 3; International Committee of the Red Cross [ICRC] & Geneva Academy, 2024, p. 8). AI-DSS should not be confused with Autonomous Weapons Systems (AWS), defined as "any weapon system with autonomy in its critical functions ... that can select ... and attack ... targets without human intervention" (Davison, 2017, p. 5). They differ in that AI-DSS can only "propose military objectives and give actionable recommendations to its (human) operators," but AWS "can engage their targets on their own" (Viveros Álvarez, 2024).

The goal of this essay is to add to the discussion around AI-DSS in two areas. First, while others have explored the potential risks and repercussions of AI-DSS implementation in military decision-making¹, few have attempted to

evaluate it quantitatively. Second, despite lively discussion on the necessity of keeping humans on the loop for AI-DSS (Scherrenburg, 2025, p. 18), much less time has gone to offering a concrete framework for determining if this human-AI relationship is working as intended that can be applied across use cases. How do we know if things are going wrong? In the absence of a timely alternative, AI-DSS could be used to evaluate its own decision-making, opening the door to a closed loop where the same system that nominated a target also certifies it was the right one. My contribution is an independent framework to evaluate AI-DSS concurrently with the system's use.

I will argue that incautious AI-DSS use in military decision-making increases civilian risk and thus must be addressed with more robust risk-management frameworks. Human approval of AI-DSS-identified military targets can be nominal only, which creates situations where AI-DSS operate functionally as fully autonomous lethal weapons systems, raising the same questions of compliance with International Humanitarian Law (IHL). This necessitates the imposition of a novel risk framework for decision-makers that protects civilians by disaggregating post-strike casualty figures by age and sex. If the profile resembles the general surrounding population, escalated review is triggered, allowing AI-DSS use to be paused when the system appears to be failing.

II. AI-DSS: AWS in sheep's clothing?

Three principles of IHL bind those who
problem without proposing a measure of it.

¹ See Bearman et al. (2026) and ICRC and Geneva Academy (2024). Both identify the verification-under-speed

would utilize AWS: distinction, proportionality, and precautions in attack (Davison, 2017, pp. 7–8). Legal analysis suggests that compliance with IHL requires limits to lawful levels of autonomy in, and a degree of human control over, weapon systems (Davison, 2017, p. 15). Thus, fully autonomous lethal weapons systems, which don't include human input in the decision-making process, cannot be used in compliance with IHL.

Unwary AI-DSS use in warfare risks a similar outcome because it can deprioritize, to the point of eliminating, human discretion in targeting decisions. Since AI-DSS can identify many more targets than human analysts, decision-makers may call more strikes because of action bias (Andersin, 2025, pp. 349–350). Overwhelmed with the quantity of identified targets and under pressure to act quickly, automation bias could then lead to insufficient target review before approval (Andersin, 2025, p. 348). Even when decision-makers genuinely want to keep humans on the loop, these compounding biases can lead to situations where “the functions of target selection and the application of force are – in effect – being carried out without any real human intervention,” resulting in a process that is analogous to fully autonomous weapons systems and similarly incompatible with IHL (ICRC & Geneva Academy, 2024, p. 17).

III. Case study: Lavender use in Gaza (worst-case scenario)

The Israel Defense Forces' (IDF) use of Lavender, an AI-DSS, is likely the premier example of insufficient human discretion causing the lines separating AI-DSS and AWS to collapse. Tal Mimran, a lecturer at Hebrew University in Jerusalem who had previous experience in Israeli military targeting operations, estimated that, prior to AI-DSS, a group of 20 officers might produce 50-100 targets in 300 days (Brumfiel, 2023). Lavender, on the other hand, generated at least 37,000 target recommendations at an early stage of the war (McKernan & Davies, 2024). One intelligence officer

stated, “I would invest 20 seconds for each target at this stage, and do dozens of them every day. I had zero added-value as a human, apart from being a stamp of approval. It saved a lot of time” (McKernan & Davies, 2024). This suggests that in these decision-making processes, humans maintain only the “illusion of control” (Boyd & Saade, 2025), the exact phenomenon described in the prior section.

Hypothetically, if Lavender was working correctly, one would expect to see an overrepresentation of military-age men in the disaggregated data; that is the pattern OHCHR verified in Gaza's earlier escalations, where men aged 20 to 39 accounted for 41 percent of fatalities in 2021. Instead, the fatalities OHCHR has verified in residential strikes since October 2023 closely resemble Gaza's age demographics, and killed women and men in roughly equal numbers², which OHCHR itself reads as raising concerns regarding compliance with the principle of distinction.

Unit 8200, the IDF's intelligence division that developed Lavender, concluded after randomly sampling and cross-checking the system's own predictions that it had achieved a 90 percent accuracy rate (McKernan & Davies, 2024). The unit that built the system is the same one that validated it, without publishing details of how the program reached its conclusions. An unverified, self-reported accuracy rate is exactly the reason an independent evaluative framework is needed. It is important to note that comments from certain Israeli officials give real reason to doubt the IDF's care in its strikes on Gaza (Associated Press & Times of Israel Staff, 2026). Whether the failure is negligent or deliberate, however, does not change what the framework's metric would detect.

2 United Nations Office of the High Commissioner for Human Rights (2024), p. 8, including note 22 on the distribution of fatalities in the 2021 and 2014 escalations, and pp. 9–11 on distinction and precautions.

IV. Methodology – Disaggregated Civilian Harm Reporting Framework

Lavender is not an American system, but it shows what the absence of an independent measure permits; the question is whether the U.S. process would detect the same pattern. DoDI 3000.17 already runs a post-strike process at the level of the individual strike. A report of civilian harm triggers an initial review that correlates the allegation against operational data; where that review finds cause, a Civilian Harm Assessment Cell conducts an assessment, and whoever conducts either step “must not have been directly involved in the event(s) being assessed” (U.S. Department of Defense [DoD], 2023, paras. 4.4.b, 4.5.a). The instruction then specifies what that assessment records: dates, locations, munitions employed, whether the targeting was deliberate or dynamic, and, of the casualties, only “a differentiation between those injured and killed” (DoD, 2023, para. 4.7.b(6)(b), p. 37). Age and sex are not required fields. However, the same paragraph directs commands to collect data for congressional reporting, and the statute requires the annual report to include, to the extent practicable, the number of men, women, and children involved³. The asymmetry is not incidental: collateral damage estimation already reasons about non-combatant occupancy before a strike, and combatant commands must incorporate demographic data into planning, while nothing requires anyone to record afterwards who was actually hit⁴. Epic Fury has laid this gap

3 National Defense Authorization Act for Fiscal Year 2018 (2017), § 1057(b)(2)(E), as amended by National Defense Authorization Act for Fiscal Year 2023 (2022), § 1056(a)(2) (C), applicable to reports submitted on or after May 1, 2024. Section 1057(d) permits a classified annex, so a demographic breakdown could lawfully exist out of public view.

4 Chairman of the Joint Chiefs of Staff (2012), pp. D-A-31–D-A-32, on the prediction of noncombatant demographics and the CDE Level 5 Population Density Reference Table; and McNerney et al. (2021), p. 56, note 54, on the requirement for combatant commands to incorporate demographic data as stated in CJCSI 3160.01C. The current revision is CJCSI 3160.01E, July 31, 2024, which is Controlled Unclassified Information and not publicly releasable.

bare: more than 120 Members of Congress asked whether Maven identified a school struck in Minab and whether a human verified the target, and the answer has not been made public (Jacobs et al., 2026, Question 3).

I propose two additions to that schema and one comparison. The assessment should record the age and sex of the casualties it counts, and measure that profile against the civilian population baseline the same instruction already requires commands to maintain during conflict⁵. Where the observed profile converges on the general population instead of skewing toward the demographic a legitimate strike would produce, that divergence should be a specified basis for the recommendation the initial review already makes on whether to proceed to a full assessment or to initiate an investigation⁶. This is a named criterion under an applicable clause, so no new machinery is needed.

Two design choices follow. Attaching the comparison this early matters because the instruction sets no deadline for completing an assessment and publishes only quarterly; a signal arriving after an operation has ended cannot pause a system still in use (DoD, 2023, para. 5.3.a). The assessment must also be conducted at strike level, since the instruction equally permits assessing “a still broader series of operations” and aggregates data “on the scale at which it was assessed,” which would defeat per-strike comparison (DoD, 2023, paras. 4.2.d, 4.7.b). Demography is also faster to verify, since combatant status can be slow to

5 U.S. Department of Defense (2023), on the requirement to identify and characterize the civilian population, including “population density, dynamics, and demographics,” as such information changes during conflict.

6 U.S. Department of Defense (2023), paras. 4.4.d(1) and (4). The initial review documents its results consistent with the data requirements of para. 4.7.b and recommends whether to proceed with an assessment or initiate an investigation.

establish against a decentralized adversary.

V. Conclusion & Limitations

In conclusion, almost everything this framework needs is already in place. DoDI 3000.17 requires an initial review whenever civilian harm is reported, requires that whoever conducts it was not involved in the strike, and ends with a recommendation on whether to escalate. What it does not require is any record of who was actually harmed, and that omission matters, because the demographic shape of the casualties is the clearest available

sign that a process has stopped distinguishing between combatants and civilians. The Department of Defense would change the instruction in two places: what an assessment must record, and the clause under which an initial review recommends escalating.

It is important to note what this framework does not claim. It measures whether a process is following its own verification steps rather than whether a strike was proportionate. It is also a lagging indicator, which I concede is a real limitation. Nevertheless, an imperfect signal that an organization watches closely is more useful than an elegant one that does not yet exist.

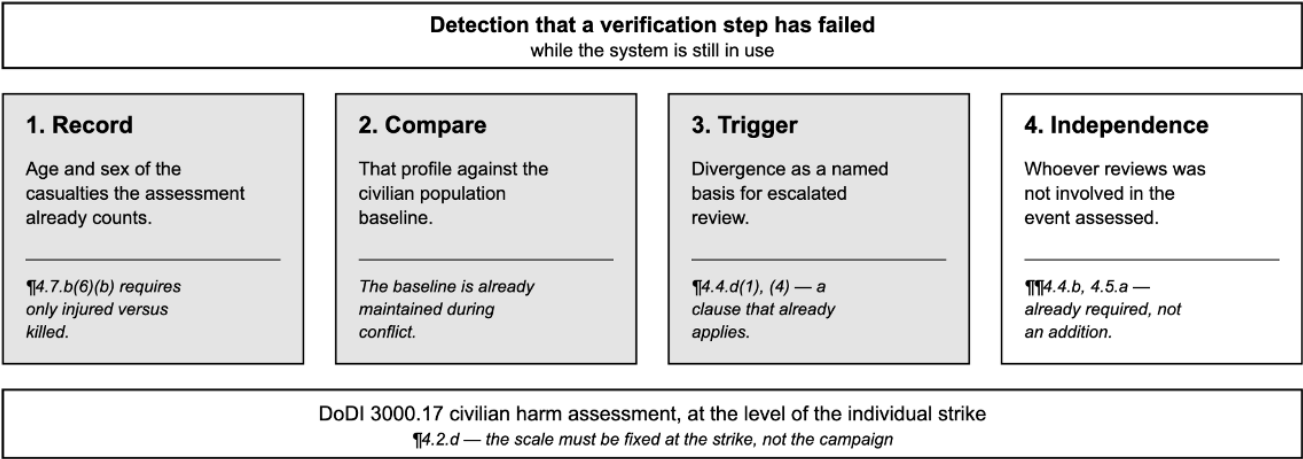


Figure 1. The four pillars of the disaggregated civilian harm reporting framework. Shading marks the three that are additions to DoDI 3000.17; independence is already required.

References

Andersin, E. (2025). The use of the “Lavender” in Gaza and the law of targeting: AI-decision support systems and facial recognition technology. *Journal of International Humanitarian Legal Studies*, 16(2), 336–370.

Associated Press, & Times of Israel Staff. (2026, August 17). Ben Gvir: We should kill “30 to 40” of Israel’s enemies in Gaza every night, “they’re not worthy of life.” *The Times of Israel*.

Bearman, T., Rosen, B., & Anderson-Samways, B. (2026, April 9). AI decision support systems: A neglected source of military AI risk [Policy memo]. Institute for AI Policy and Strategy.

Boyd, B. L., & Saade, T. (2025, April 1). Human cognitive autonomy: Keeping people on the loop in AI-driven warfare. *The Republic*.

Brumfiel, G. (2023, December 14). Israel is using an AI system to find targets in Gaza. Experts say it’s just the start. *NPR*.

Chairman of the Joint Chiefs of Staff. (2012, October 12). No-strike and the collateral damage estimation methodology (CJCSI

3160.01A). U.S. Department of Defense.

Davison, N. (2017). A legal perspective: Autonomous weapon systems under international humanitarian law. In Perspectives on lethal autonomous weapon systems (UNODA Occasional Papers No. 30, pp. 5–18). United Nations Office for Disarmament Affairs.

Freedberg, S. J., Jr. (2026, May 12). “Insatiable appetite” for AI: Maven usage surged for strikes on Iran, Pentagon AI chief says. Breaking Defense.

Holland Michel, A. (2024). Decisions, decisions, decisions: Computation and artificial intelligence in military decision-making. International Committee of the Red Cross.

International Committee of the Red Cross, & Geneva Academy of International Humanitarian Law and Human Rights. (2024, March). Expert consultation report on artificial intelligence and related technologies in military decision-making on the use of force in armed conflicts: Current developments and potential implications. International Committee of the Red Cross.

Jacobs, S., et al. (2026, March 12). Letter to Secretary of Defense Pete Hegseth on civilian casualties in Operation Epic Fury [Letter]. U.S. House of Representatives.

McKernan, B., & Davies, H. (2024, April 3). “The machine did it coldly”: Israel used AI to identify 37,000 Hamas targets. The Guardian.

McNerney, M. J., Tarini, G., Sudkamp, K. M., Lewis, L., Gris  , M., & Moore, P. (2021). U.S. Department of Defense civilian casualty policies and procedures: An independent assessment. RAND Corporation.

National Defense Authorization Act for Fiscal Year 2018, Pub. L. No. 115-91,    1057 (2017), as amended by National Defense Authorization Act for Fiscal Year 2023, Pub. L. No. 117-263,   

1056 (2022).

Scherrenburg, M. (2025, August 4). Human oversight in AI-driven defence: At what positions do we need the human in the loop? NATO Command and Control Centre of Excellence.

United Nations Office of the High Commissioner for Human Rights. (2024, November 6). Six-month update report on the human rights situation in Gaza: 1 November 2023 to 30 April 2024.

U.S. Department of Defense. (2023, December 21). Civilian harm mitigation and response (DoD Instruction 3000.17).

Viveros   lvarez, J. S. (2024, September 4). The risks and inefficacies of AI systems in military targeting support. Humanitarian Law & Policy. International Committee of the Red Cross.

AI Supported Professional Judgment: Explaining Situated Human-Machine Decision-making

Timothy Jones

Independent Researcher

Corresponding author: timothy.d.jones@northumbria.ac.uk

Introduction

AI decision-support systems analyse information and generate courses of action. Military planning involves multiple actors, incomplete information, changing conditions and demands exceeding unaided human processing capacity (Kase et al., 2022; Probasco et al., 2024). Military scholarship raises concerns about automation bias, trust, responsibility and the limits of formal human oversight (Johnson, 2023; Erskine & Miller, 2025). In sport, AI supports tactical analysis, injury prediction and performance modelling. However, research has focused primarily on the performance of these systems, with less attention to how professionals interpret and use outputs (Elstak et al., 2024; Teixeira et al., 2025). Both settings raise questions about AI's influence on professional judgement, despite different consequences and accountability.

AI-generated outputs may reinforce existing interpretations, introduce alternatives, prompt reconsideration or be discounted (Fahy et al., 2026; Liang et al., 2026; Shandler et al., 2026; Wang et al., 2024). Professional expertise, uncertainty, trust and organisational expectations may shape these responses (Martins et al., 2026; Nitzl et al., 2025; Lamb et al., 2026; Salatino et al., 2026). This paper therefore asks: under what conditions, and through what mechanisms might AI-supported decision environments enhance, constrain, or redirect professional judgement?

To address this question, the paper adopts a Critical Realism (CR) perspective, which understands outcomes as contingent on the interaction of causal powers and conditions in open systems (Bhaskar, 1998; Danermark et

al., 2002). The comparison examines how task demands, professional expertise, the timing of AI advice and organisational arrangements relate to reported responses (Dugle et al., 2020). Abductive and retroductive reasoning is used to develop provisional explanations for this variation, without claiming to verify underlying mechanisms (Meyer & Lunnay, 2013).

Method

Data sources and selection

This paper uses a CR-informed comparative narrative synthesis of eleven empirical studies: four from sport and seven from military settings. Studies were selected purposively for evidence of how professionals encounter, interpret or act upon AI-supported information. Each study is charted regarding its setting and design, the AI-supported task, evidence of human engagement, and the principal finding and limitations (Table 1).

The reviewed studies include interviews, experiments, simulations and expert evaluations. Findings were synthesised narratively, without statistical pooling or line-by-line coding (Rodgers et al., 2009). Findings were examined within each domain, then the evidence contrasted across seven emerging analytical issues: what becomes salient, task structure, acceptance or resistance, expertise, organisational mediation, repeated interaction and human control (Table 2). These issues organised the comparison, without imposing predetermined theoretical codes. Abduction then involved reconsidering contrasting responses, including compliance (Salatino et al., 2026), aversion (Shandler et al., 2026) and reconsideration (Liang et al., 2026). Retroduc-

tion asked what conditions might explain these differences, including uncertainty, prior judgement, professional expertise and organisational arrangements (Danermark et al., 2002; Meyer & Lunnay, 2013). The resulting propositions were considered against supporting and contrasting findings, consistent with CR comparative reasoning (Ackroyd & Karlsson, 2014; Williams & Karahanna, 2013).

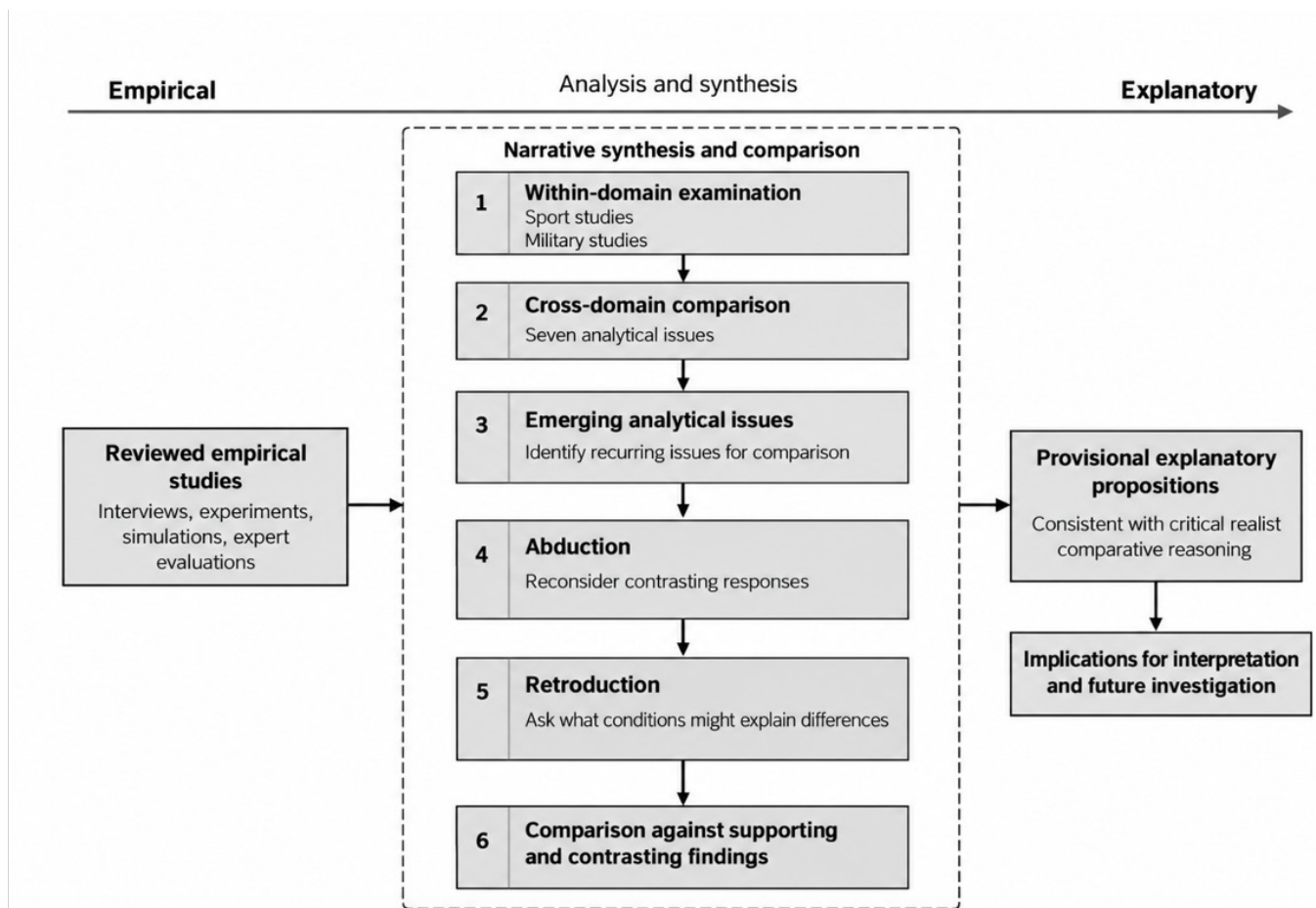


Figure 1. Process of comparative narrative synthesis and explanatory development. Adaptation of Radulescu and Vessey (2009).

Findings and discussion

Timing of AI advice

AI advice can enter decision-making at different stages. In military planning, LLM outputs sometimes reinforced officers' initial interpretations during problem framing (Fahy et al., 2026).

Other military studies examined advice presented immediately before a decision (Salatino et al., 2026) or after an initial judgement (Tossell et al., 2025). Similarly, in simulated basketball scenarios, coaches sometimes revised their initial decisions after receiving AI recommendations (Liang et al., 2026).

Advice received while a problem is being framed may shape which alternatives are considered (Agudo et al., 2024) whereas later advice can be weighed against an existing judgement (Tossell et al., 2025). Indeed, the order of advice and evidence has been found to influence decisions (Aiyer & Yeung, 2025). Although timing was not isolated as a causal factor in the reviewed sport and military studies, evidence suggests that it may condition AI's influence. This matters for human control because AI may shape how a problem is understood even when the final decision remains human (Johnson, 2023; Erskine & Miller, 2025).

Problem structure and uncertainty

The contribution of AI may depend, in part, on the characteristics of the problem being addressed. For example, 'TacticAI' operates within relatively bounded soccer corner-kick situations (Wang et al., 2024). By contrast, injury-risk decisions require practitioners to interpret predictions alongside clinical assessments, athlete histories and competing demands (Martins et al., 2026). This contrast suggests that the value of AI-generated information may depend on how readily the problem can be represented computationally and how much contextual interpretation remains necessary.

In military settings, Nitzl et al. (2025) found that AI improved performance under time pressure, although this advantage weakened when intelligence became ambiguous or contradictory. One possible explanation is that AI can contribute more directly when information and decision requirements are clearly specified, while uncertainty increases the importance of contextual professional expertise (Kase et al., 2022; Probasco et al., 2024). From a CR perspective, problem structure may therefore condition how AI capabilities are realised in practice, although the available evidence cannot establish this relationship independently of other contextual influences.

Responses to AI advice

Military studies suggest that neither compliance nor resistance is an inevitable response to AI. Salatino et al. (2026) found that military participants were more likely to follow incorrect AI recommendations in simulated drone-strike decisions. In contrast, Shandler et al. (2026) found that military personnel were less likely to approve AI-sourced targeting recommendations than comparable human-sourced recommendations, particularly when anticipated collateral damage was high. Dobos et al. (2026), meanwhile, found no significant effect of AI-like labels. These contrasting findings suggest that responses cannot be explained by generalised trust alone. The perceived consequences of a decision, the nature of the AI input, and opportunities to assess its credibility may help explain why advice is accepted or resisted.

Evidence from sport research provides further insight into how professionals evaluate and respond to AI-generated information. Martins et al. (2026) described practitioners considering AI-informed injury-risk predictions alongside clinical expertise and through multidisciplinary discussion. Liang et al. (2026) found that AI information sometimes prompted coaches to revise their initial decision. These findings can be considered alongside Shandler et al. (2026), who found that military per-

sonnel were less likely to approve AI-sourced recommendations than comparable human-sourced recommendations, although this difference was no longer evident when additional explanatory information was provided. While the studies examine different professional responses, they raise questions about how existing expertise, opportunities to evaluate AI information and the way recommendations are presented might influence professional judgement. Digital literacy and organisational support for challenging AI advice may further influence these interactions (Ghorbel et al., 2026; Heng, 2025). However, the different research designs and decision contexts limit what can be concluded about the underlying causal processes.

Organisational mediation and learning

AI advice is interpreted within existing professional relationships. Coaches anticipated analysts mediating outputs (Aarons et al., 2024), while injury-risk predictions were discussed among coaches and performance staff (Martins et al., 2026). Similarly, military intelligence adoption reflected team practices and organisational incentives (Lamb et al., 2026). Müller et al. (2025) argue that how systems enter established decision processes matters. Across settings, one possible explanation is that organisational arrangements influence whose expertise shapes AI advice and whether outputs are questioned.

These arrangements may evolve through use. Lamb et al. (2026) found analysts increasingly treating an imperfect machine-vision system as a tool after learning its limitations. Repeated engagement might therefore change how subsequent outputs are evaluated. Yet evidence of lasting organisational transformation remains limited (Erskine & Davis, 2025).

Contribution to the Future of Decision-Making project

This paper contributes to the Future of Decision-Making project through a compara-

tive account of how AI-supported information can influence professional judgement. It addresses human-machine collaboration and the implications of AI for expertise, trust and responsibility, connecting particularly with the project's Organisations, Governance and Capability Development themes.

Organisational and process design should consider when advice appears and who interprets it. Governance questions include how to communicate uncertainty, preserve opportunities for independent judgement and enable professionals to challenge recommendations. Capability development concerns whether professionals can assess AI information alongside their own expertise and other evidence. The synthesis identifies these issues as priorities for investigation; it does not determine which interventions are effective.

Future research

Future studies should examine AI-supported judgement as a situated process unfolding over time. The Critical Decision Method (Klein et al., 1989) could reconstruct the cues and interpretations shaping individual decisions. Abbott (2001) highlights the significance of event sequence, suggesting comparisons of advice introduced before and after initial judgements. Longitudinal research informed by Archer (1995) could examine how existing organisational arrangements condition AI use and whether repeated interaction reproduces or changes those arrangements. Such work would help test the provisional explanations while recognising that timing and organisational context are potentially consequential, rather than neutral backdrops.

References

Aarons, M. F., Vickery, W., Bruce, L., Young, C. M., & Dwyer, D. B. (2024). Barriers to coach decision-making during Australian football matches and how it can be supported by artificial intelligence. *International Journal*

of Sports Science & Coaching, 19(1), 41–52.
<https://doi.org/10.1177/17479541231206682>

Abbott, A. (2001). *Time matters: On theory and method*. University of Chicago Press.

Ackroyd, S., & Karlsson, J. C. (2014). Critical realism, research techniques, and research designs. In P. K. Edwards, J. O'Mahoney, & S. Vincent (Eds.), *Studying organizations using critical realism: A practical guide* (pp. 21–45). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199665525.003.0002>

Agudo, U., Liberal, K. G., Arrese, M., & Matute, H. (2024). The impact of AI errors in a human-in-the-loop process. *Cognitive Research: Principles and Implications*, 9, Article 1. <https://doi.org/10.1186/s41235-023-00529-3>

Aiyer, S., & Yeung, N. (2025). The impact of the timing of advice on its utilization. *Journal of Behavioral Decision Making*, 38(3), Article e70021. <https://doi.org/10.1002/bdm.70021>

Archer, M. S. (1995). *Realist social theory: The morphogenetic approach*. Cambridge University Press.

Bhaskar, R. (1998). *The possibility of naturalism: A philosophical critique of the contemporary human sciences* (3rd ed.). Routledge.

Danermark, B., Ekström, M., Jakobsen, L., & Karlsson, J. C. (2002). *Explaining society: Critical realism in the social sciences*. Routledge.

Dobos, N., Boshuijzen-van Burken, C., Baker, D.-P., Ghasrihouzani, M., Hene Kankanamge, E., Huybers, T., Molloy, O., Plested, J., & Veda, A. (2026). Do AI-enabled decision-support systems improve ethical choices on the battlefield? *Australian Army Journal*, 22(2). <https://doi.org/10.61451/220201>

Dugle, G., Wulifan, J. K., Tanyeh, J. P., &

Quentin, W. (2020). A critical realist synthesis of cross-disciplinary health policy and systems research: Defining characteristic features, developing an evaluation framework and identifying challenges. *Health Research Policy and Systems*, 18, Article 79. <https://doi.org/10.1186/s12961-020-00556-2>

Elstak, I., Salmon, P., & McLean, S. (2024). Artificial intelligence applications in the football codes: A systematic review. *Journal of Sports Sciences*, 42(13), 1184–1199. <https://doi.org/10.1080/02640414.2024.2383065>

Erskine, T., & Davis, J. L. (2025). “Borgs in the org” and the decision to wage war: The impact of AI on institutional learning and the exercise of restraint. *Cambridge Forum on AI: Law and Governance*, 1, Article e45. <https://doi.org/10.1017/cfl.2025.10037>

Erskine, T., & Miller, S. E. (2025). Computer says, “war”: AI and resort-to-force decision making in a context of rapid change and global uncertainty. *Cambridge Forum on AI: Law and Governance*, 1, Article e47. <https://doi.org/10.1017/cfl.2025.10041>

Fahy, C., McDonald, N., Keenan, K., & Corrigan, S. (2026). An exploratory study into the relationship between sycophantic behavior in large language models and military decision makers engaged in naturalistic decision making. *AI Magazine*, 47(3), Article e70073. <https://doi.org/10.1002/aaai.70073>

Ghorbel, A., Yaakoubi, M., Trabelsi, O., Elkholi, S. M., Kammoun, M. M., Masmoudi, L., & Gharbil, A. (2026). Digital literacy shapes AI chatbot adoption, barriers, and performance outcomes in football coaches: A multi-method analytical study. *Scientific Reports*. <https://doi.org/10.1038/s41598-026-72410-4>

Heng, Y.-K. (2025). Upskilling human actors against AI automation bias in strategic decision making on the resort to force. *Cambridge Forum on AI: Law*

and Governance, 1, Article e53. <https://doi.org/10.1017/cfl.2025.10029>

Johnson, J. (2023). The AI commander problem: Ethical, political, and psychological dilemmas of human-machine interactions in AI-enabled warfare. *Journal of Military Ethics*, 21(3–4), 246–271. <https://doi.org/10.1080/15027570.2023.2175887>

Kase, S. E., Hung, C. P., Krayzman, T., Hare, J. Z., Rinderspacher, B. C., & Su, S. M. (2022). The future of collaborative human-artificial intelligence decision-making for mission planning. *Frontiers in Psychology*, 13, Article 850628. <https://doi.org/10.3389/fpsyg.2022.850628>

Klein, G., Calderwood, R., & MacGregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(3), 462–472. <https://doi.org/10.1109/21.31053>

Lamb, S. C., Byrne, M., Norman, T. J., Ramchurn, S. D., & Naiseh, M. (2026). What's informing trust and adoption? Can current cognitive models support expert adoption of AI? *Journal of Cognitive Engineering and Decision Making*, 20(2), 158–175. <https://doi.org/10.1177/15553434261430550>

Liang, Y., Guo, X., Zhang, J., & Du, X. (2026). NeuroPlayNet: A multimodal AI framework for real-time cognitive-aware strategy optimization in professional basketball. *Scientific Reports*, 16, Article 15085. <https://doi.org/10.1038/s41598-026-41140-y>

Martins, F., Gouveia, É. R., Przednowek, K., Sousa, H., & Sarmiento, H. (2026). Beyond the algorithm: A qualitative study of expert perceptions on the impact of injury prediction models in men's professional football. *Journal of Science and Medicine in Sport*. Advance online publication. <https://doi.org/10.1016/j.jsams.2026.09.010>

Meyer, S. B., & Lunnay, B. (2013). The application of abductive and retroductive inference for the design and analysis of theory-driven sociological research. *Sociological Research Online*, 18(1), 86–96. <https://doi.org/10.5153/sro.2819>

Müller, D., Chiodo, M., & Sienknecht, M. (2025). Integrators at war: Mediating in AI-assisted resort-to-force decisions. *Cambridge Forum on AI: Law and Governance*, 1, Article e51. <https://doi.org/10.1017/cfl.2025.10030>

Nitzl, C., Cyran, A., Krstanovic, S., & Borghoff, U. M. (2025). The use of artificial intelligence in military intelligence: An experimental investigation of added value in the analysis process. *Frontiers in Human Dynamics*, 7, Article 1540450. <https://doi.org/10.3389/fhumd.2025.1540450>

Probasco, E., Burtell, M., Toner, H., & Rudner, T. G. J. (2024). Not oracles of the battlefield: Safety considerations for AI-based military decision support systems. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1), 1157–1165. <https://doi.org/10.1609/aies.v7i1.31712>

Radulescu, C., & Vessey, I. (2009). Methodology in critical realist research: The mediating role of domain specific theory. In *Proceedings of the 15th Americas Conference on Information Systems* (Paper 433). <https://aisel.aisnet.org/amcis2009/433/>

Rodgers, M., Sowden, A., Petticrew, M., Arai, L., Roberts, H., Britten, N., & Popay, J. (2009). Testing methodological guidance on the conduct of narrative synthesis in systematic reviews: Effectiveness of interventions to promote smoke alarm ownership and function. *Evaluation*, 15(1), 49–73. <https://doi.org/10.1177/1356389008097871>

Salatino, A., Prével, A., Caspar, E. A., & Lo Bue, S. (2026). The interplay of time pressure and AI assistance reveals moral responsibility recalibration in high-stakes decisions. *Scientific*

Reports. <https://doi.org/10.1038/s41598-026-69263-2>

Shandler, R., Gross, M. L., & Shereshevsky, Y. (2026). Black box warfare: Human judgment and military decision-making in the age of AI. *Journal of Conflict Resolution*. Advance online publication. <https://doi.org/10.1177/00220027261463443>

Teixeira, J. E., Maio, E., Afonso, P., Encarnação, S., Machado, G. F., Morgans, R., Barbosa, T. M., Monteiro, A. M., Forte, P., Ferraz, R., & Branquinho, L. (2025). Mapping football tactical behavior and collective dynamics with artificial intelligence: A systematic review. *Frontiers in Sports and Active Living*, 7, Article 1569155. <https://doi.org/10.3389/fspor.2025.1569155>

Tossell, C. C., Kuennen, C., Momen, A., Funke, G., Tolston, M., & De Visser, E. J. (2025). Robots in the moral loop: A field study of AI advisors in ethical military decision-making. *Frontiers in Artificial Intelligence*, 8, Article 1694772. <https://doi.org/10.3389/frai.2025.1694772>

Wang, Z., Veličković, P., Hennes, D., Tomašev, N., Prince, L., Kaisers, M., Bachrach, Y., Elie, R., Wenliang, L. K., Piccinini, F., Spearman, W., Graham, I., Connor, J., Yang, Y., Recasens, A., Khan, M., Beauguerlange, N., Sprechmann, P., Moreno, P., ... Tuyls, K. (2024). TacticAI: An AI assistant for football tactics. *Nature Communications*, 15, Article 1906. <https://doi.org/10.1038/s41467-024-45965-x>

Williams, C. K., & Karahanna, E. (2013). Causal explanation in the coordinating process: A critical realist case study of federated IT governance structures. *MIS Quarterly*, 37(3), 933–964. <https://doi.org/10.25300/MISQ/2013/37.3.12>

Appendix

Table 1. Reviewed Empirical Studies

Study	Domain / participants	AI-supported task	Human-AI evidence	Principal finding	Key limitation
Aarons et al. (2024)	Sport; 6 elite Australian football head coaches	Prototype in-match DSS: predictions, suggestions, simulation	Interviews about real decision constraints plus prototype interaction	Coaches envisaged AI testing hypotheses, prompting discussion and exploring options, usually mediated by analysts	Prospective/anticipated use rather than observed live AI-supported decisions
Martins et al. (2026)	Sport; 10 professional-football performance staff	Injury-risk prediction	Interviews on integration of prediction models into practice	Outputs required clinical judgement, longitudinal knowledge, organisational context and negotiation	Practitioner accounts rather than direct observation of decisions
Wang et al. (2024)	Sport; 5 Liverpool FC domain experts	TacticAI prediction, retrieval and generation of corner-kick alternatives	Expert evaluation of 50 AI-generated tactical refinements	AI expanded the option set in a bounded tactical problem; 45/50 recommendations judged favourable on average	Evaluates recommendations, not subsequent coaching decisions
Liang et al. (2026)	Sport; professional basketball coaches/analysts in simulation	Multimodal real-time recommendations	Initial decisions compared with decisions after recommendations	27 decision changes reported across 54 simulations; explanations appeared important to reconsideration	Proof-of-concept; simulated setting and substantial synthetic data
Shandler et al. (2026)	Military; 2,015 Israeli military personnel	High-fidelity targeting DSS	Two experiments manipulating AI/human source and interface information	Strong algorithmic aversion, especially with higher collateral-damage stakes; explanations reduced aversion	Experimental scenarios rather than operational combat

Study	Domain / participants	AI-supported task	Human-AI evidence	Principal finding	Key limitation
Lamb et al. (2026)	Military; 12 intelligence analysts pre-use, 9 longitudinal completers	Machine-vision object detection in analyst workflows	Naturalistic longitudinal adoption study	Reliance changed through experience; system boundaries and configurability became more important; perceived role shifted from adviser toward tool	Small specialist sample
Fahy et al. (2026)	Military; 30 Captains, ≥6 years' service	LLM-supported tactical decision game	Experimental problem framing and COA generation	LLM language could reinforce initial framing and facilitate progression without necessarily prompting critical reconsideration	Exploratory design; AI prompt behaviour itself was imperfect
Nitzl et al. (2025)	Military; 29 active-duty soldiers + 7 expert analysts	AI search, summarisation and entity recognition	AI versus non-AI experimental comparison	AI improved overall performance under time pressure; advantage weakened with ambiguous/contradictory material	Small sample; corpus version is manuscript/preprint
Salatino et al. (2026)	Military; 36 cadets/officers	AI advice in simulated drone-strike decisions	Within-subject AI/no-AI and time-pressure manipulation	AI increased compliance with clearly wrong recommendations and reduced felt responsibility; time pressure did not simply amplify reliance	Simulated moral decisions

Study	Domain / participants	AI-supported task	Human-AI evidence	Principal finding	Key limitation
Tossell et al. (2025)	Military; 62 US military cadets	Human/robotic moral adviser	Three-month teaming followed by sequential recommendation/revision task	Around half revised an initial judgement; many shifted again after opposing advice	Cadets rather than experienced commanders; ethical vignette
Dobos et al. (2026)	Military; 81 ADFA cadets/trainee officers	AI-like visual classification in shoot/no-shoot scenarios	Behavioural experiment across 12 video scenarios	AI labels produced no statistically significant effect on decision correctness	Trainees; participants knew labels could be wrong

Table 2. Comparative analysis findings

Analytical issue	Sport evidence	Military evidence	Comparative interpretation
What becomes salient	TacticAI introduces alternative tactical configurations; Aarons' coaches envisage predictions testing existing hypotheses.	Fahy shows LLM output can amplify the user's initial framing.	AI can influence judgement before the final recommendation by changing the informational field.
Structure of the problem	TacticAI operates in bounded corner-kick situations; Martins shows injury decisions require contextual integration.	Nitzl finds strongest AI gains in information-processing tasks, with reduced advantage as ambiguity increases.	The usefulness of AI appears contingent on how computationally tractable the underlying problem is.
Acceptance versus resistance	Martins describes contextualisation and negotiation; Aarons anticipates selective use.	Salatino shows over-reliance; Shandler shows aversion; Dobos finds little measurable influence.	Machine outputs receive different epistemic weight under different conditions.
Role of expertise	Coaches and performance staff incorporate history, relationships, tactical intentions and individual knowledge.	Lamb shows experts learn system boundaries; Fahy shows prior framing shapes AI interaction.	Professionals actively mediate AI rather than merely receive its recommendations.
Organisational mediation	Analysts, coaches, medical/performance staff and athletes mediate outputs.	Intelligence teams, command structures and organisational incentives affect adoption.	Human-machine decision-making is often a socio-technical system, not a human-machine dyad.
Time and repeated interaction	Very limited direct evidence.	Lamb demonstrates changing trust and role conceptions through use; Tossell shows sequential recalibration.	Human-AI relations may be recursive, but longitudinal evidence remains limited.
Human control	Formal coach authority remains, but AI can shape options and evidence before action.	Final authority may remain human while AI affects framing, reliance and responsibility.	'Human in the loop' does not by itself establish meaningful epistemic control.

The Aid Sector at Machine Speed: AI and the Future of Humanitarian Decision-making

Rachel A. George, PhD

Stanford University,

Corresponding author: rageorge@stanford.edu

Aid Under Strain

The international aid system is facing a crisis of unprecedented strain. In 2025, the United States cut bilateral foreign assistance by 56.9 percent following USAID's dismantling, and official development assistance from OECD DAC members fell 23.1 percent, the largest single-year decline on record (OECD, 2026).

Into this moment of aid contraction has arrived artificial intelligence. Within roughly thirty-six months of ChatGPT's public release, AI tools have moved from experimental pilots to routine workflows in ministries, donor agencies, and front-line NGOs — driven less by organizational planning than by peer pressure and individual initiative¹. Early evidence of promise exists: a randomized trial of Microsoft Copilot in Edo State, Nigeria produced learning gains equivalent to 1.5 to 2 years of standard schooling in six weeks (De Simone et al., 2025); Zipline's drone network cut blood product delivery times and wastage at remote Rwandan health facilities (Nisingizwe et al., 2022). The challenge is institutionalizing and scaling these results.

This essay argues that AI's role in decision-making for foreign assistance and humanitarian aid will shape the next frontier for aid effectiveness, perhaps more than financial factors. Just as AI's

military use demands urgent attention, so does its use in the systems built to prevent

¹ Interviews with humanitarian and development practitioners in donor agencies in the US, UK, and Norway, and recipients in Sudan and the Philippines, March-May 2026.

and respond to conflict.

The central risk is an automation trap: governments and agencies, facing shrinking budgets and pressure to demonstrate efficiency, are layering AI onto systems that were already struggling. Rather than a tool for reform, AI risks becoming a response to institutional decline — compressing decision-making, displacing expertise, and encoding existing failures at greater speed and scale. Research links effective aid to reduced poverty and conflict risk (Collier & Hoeffler, 1998; Stewart, 2004; Homer-Dixon, 1994), so automated failures reach beyond recipient communities. Drawing on practitioner interviews, this essay examines the opportunities and risks AI introduces into aid decision-making and the reforms organizations must pursue before automating.

Aid's Automation Trap

The automation trap is the tendency of institutions under fiscal or political pressure to deploy AI to optimize existing administrative structures rather than redesign flawed decision architectures. In development, AI would be layered onto the donor-driven, country-office, vertical-programming architecture as an efficiency play, consistent with DiMaggio and Powell's (1983) isomorphic pressures, whereby organizations under uncertainty adopt practices that signal legitimacy to funders and peers rather than maximize mission effectiveness.

Beyond well-documented linguistic imbalance and substantive bias in many models against the Global South (Buolamwini

& Gebru, 2018; Sambasivan et al., 2021), AI embeds assumptions about what decisions need to be made, who is authorized to make them, how quality is measured, and who bears the consequences of error (Acemoglu & Johnson, 2023; Crawford, 2021).

In aid decision-making, institutions might deploy AI for efficiency gains most legible to those controlling resources, especially in monitoring and evaluation (M&E) systems, the evidence chain between field operations and donor capitals. Automating reporting and monitoring could make this architecture run faster without making it work better, and AI's potential to amplify recipient voices at scale (Jean et al., 2016; World Bank, 2021) remains underutilized when institutions skip the redesign step.

Organizational Case: AI-Enabled Results Monitoring

Development agencies have long operated under what Michael Power (1997) terms “audit society” conditions: facing legitimacy pressure, they generate elaborate evidence of activity while true impact remains elusive. Program officers log outputs against donor-specified indicators, which flow upward into portfolio-level reports that unlock funding tranches.

AI can serve within this infrastructure as an accelerant, aggregating data across hundreds of sub-grantees, flagging anomalies, generating narrative summaries from raw output logs, and compressing production timelines for donor reports. In fragile and conflict-affected states, M&E systems govern resource allocation for stabilization and conflict prevention, where development most directly intersects with international security. When these systems optimize for donor legibility rather than operational accuracy, the result can be misallocated stabilization funds, delayed humanitarian response, and interventions calibrated for reporting metrics rather than local politi-

cal realities.

AI systems trained on standardized data risk simplifying complex local realities faster but no more accurately: what appears as data quality improvement may instead be legibility improvement, where the aid system learns to see more of what it can already measure, not more of what matters. Sector-wide normalization of AI adoption compounds this risk, a pattern Madianou (2019) links to entrenched institutional power asymmetries.

Consequences for Aid Decision-Making

The automation trap consequences for aid decision-making, with broader implications for mission-driven organizations.

First, it accelerates decisions without necessarily improving them. Integrated into flawed decision processes, AI can produce faster decisions with greater epistemic authority, even when poorly calibrated to local realities. Algorithmic needs assessments may carry more institutional weight than field officers' contextual judgment, despite being less sensitive to political and social conditions (Mohamed, Png, & Isaac, 2020; Birhane, 2020).

Second, AI can restructure accountability in ways that privilege institutional legibility over beneficiary visibility. Organizations that automate reporting and monitoring improve transparency to donors while reducing the visibility of decisions to affected communities. In Ibaji, Nigeria, GiveDirectly's AI-triggered flood cash program narrowed roughly 39,000 applicants to about 4,600 recipients through automated and call-center verification. An investigation found some of the area's poorest residents excluded over name mismatches and phone-charging costs, some without explanation (Ayeni, 2026); GiveDirectly calls its design choices deliberate tradeoffs favoring speed and coverage. Because beneficiaries are far removed from formal accountability mechanisms, such exclusions can be hard to

detect.

Third, automation can hollow out institutional expertise. Organizations that delegate cognitive work to AI without preserving the capacity to interrogate outputs risk eroding the contextual knowledge required to recognize when systems are wrong. Field officers become data-entry points and the expertise needed to interpret local conditions disappears from downstream allocation decisions (Acemoglu & Johnson, 2023). Interviewees consistently identified this erosion of field-level discretion as one of the least visible but most consequential risks of unreflective AI adoption².

Governing the Transition: Three Principles

Avoiding the automation trap and reaping AI's rewards requires treating adoption as an opportunity for institutional redesign.

Redesign before automating. Aid organizations that map decision-making structures before deploying AI can better shape outcomes, building systems that collect beneficiary feedback upstream and use AI to route funding more directly to local actors and broaden the voices reaching decision-makers. In practice, that means a written decision map before procurement: for each decision the tool touches, who decides, on what evidence, against what standard, and who bears the cost of error. Fajrillah et al. (2026) show the sequence: before designing any machine-learning integration, they documented the warning signs Indonesian fishing communities already trust, though no AI system has yet been built or tested.

Build accountability mechanisms that measure outcomes, not process efficiency

² This risk was particularly prominent among Canadian, Norwegian and British practitioners concerned about AI's impacts on localization commitments.

cy. Procurement standards and evaluation frameworks should require evidence that AI-enabled systems improve decision quality against mission objectives, not merely speed or cost per output. Building on the OECD-DAC's emerging guidance (OECD, 2026), funders should require, for AI-enabled targeting or verification, exclusion counts by stage with reasons, the share of payments delivered late, human review of high-stakes allocations, and a reasoned notice with human appeal for each exclusion. Since evaluators remain responsible for explaining AI limits (Jacob, 2026), AI-generated syntheses should reach donor reports only after a named evaluator validates them. Contracts should settle platform and data ownership, including export and exit rights, since the risk lies in AI run on infrastructure an organization cannot control (Vinck & Pham, 2026).

Invest in institutional capacity to interrogate AI outputs. Organizations that deploy AI systems they cannot explain to affected stakeholders must build the analytical and human capacity to do so, preserving domain expertise and cultures that reward catching machine errors. AI can bring valuable evidence and analysis to aid systems, but requisite expertise in technological tools and verification infrastructure is needed in the decision chain for allocation, program design, and evaluation, particularly where decisions are consequential and contestable (George, 2026). Few organizations can audit models alone, so donors should fund shared assessment capacity pooled across agencies and regions, and field staff should hold real authority to override automated flags (Vinck & Pham, 2026).

Conclusion

The aid automation trap is not a technological inevitability. Early evidence of AI's effectiveness in foreign assistance systems offers real promise that remains unrealized at scale without organizational redesign. A sector that automates its way into deeper dysfunction removes one of the primary

non-military instruments through which states manage the conditions that generate security threats, just as strategic competition over development influence makes effective aid more consequential. Organizations can use this window to redesign decision-making architectures, building AI into structures that route resources closer to beneficiaries and preserve the human judgment that makes accountability real. Those that use it to automate existing processes will move faster toward familiar failures.

References

- Acemoglu, D., & Johnson, S. (2023). Power and progress: Our thousand-year struggle over technology and prosperity. PublicAffairs.
- Ayeni, D. (2026, August 31). GiveDirectly's AI experiment left vulnerable flood victims behind in Nigeria. *The New Humanitarian*. <https://www.thenewhumanitarian.org/investigations/2026/08/31/givedirectlys-ai-experiment-left-flood-victims-behind-nigeria>
- Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPTed*, 17(2), 389–409.
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 77–91.
- Collier, P., & Hoeffler, A. (1998). On economic causes of civil war. *Oxford Economic Papers*, 50(4), 563–573.
- Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
- De Simone, E., et al. (2025). From chalkboards to chatbots: Evaluating the impact of generative AI on learning outcomes in Nigeria. World Bank Group.
- DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, 48(2), 147–160.
- Fajrillah, A. A. N., Hartanto, R., & Nugroho, L. (2026). From narrative to machine-readable logic: Formalising and validating local indigenous knowledge for rural early warning systems. *Jàmbá: Journal of Disaster Risk Studies*, 18(1), Article a2036. <https://doi.org/10.4102/jamba.v18i1.2036>
- George, R. A. (2026). *The Next National Security Challenge is Research Integrity*. Tech Policy Press.
- Homer-Dixon, T. F. (1994). Environmental scarcities and violent conflict: Evidence from cases. *International Security*, 19(1), 5–40.
- Jacob, S. (2026). Beyond the hype: Uncovering pitfalls in the use of AI in evaluation. *Canadian Journal of Program Evaluation*, 40(1), 1–17. <https://doi.org/10.18357/cjpe.2026.40.1.1210>
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790.
- Madianou, M. (2019). Technocolonialism: Digital innovation and data practices in the humanitarian response to refugee crises. *Social Media + Society*, 5(3), 1–13.
- Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4), 659–684.
- Nisingizwe, M. P., et al. (2022). Effect of unmanned aerial vehicle (drone) delivery on blood product delivery time and wastage in Rwanda: A retrospective, cross-sectional study and time series analysis. *The Lancet Global Health*, 10(4), e564–e569.
- OECD. (2026). *Development co-operation report 2026*. OECD Publishing.

Power, M. (1997). *The audit society: Rituals of verification*. Oxford University Press.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. (2021). “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Article 39, 1–15.

Stewart, F. (2004). Development and security. *Conflict, Security & Development*, 4(3), 261–288.

Vinck, P., & Pham, P. N. (2026, July 29). Humanitarian artificial intelligence begins before the model: Lessons from KoboToolbox for responsible digital commons. *Humanitarian Alternatives*, Issue 32. <https://www.alternatives-humanitaires.org/en/2026/07/29/humanitarian-artificial-intelligence-begins-before-the-model-lessons-from-kobotoolbox-for-responsible-digital-commons/>

World Bank. (2021). *World development report 2021: Data for better lives*. World Bank.

Self-determined nudging and boosting cognitive competences to improve decision-making in the security-related sector: Exploratory sketch towards a CoBoNu framework

Anja Dannewitz¹, Frank Flemisch^{1,2},

¹Fraunhofer FKIE, Fraunhoferstraße 20, 53343 Wachtberg, Germany

²Institut für Arbeitswissenschaft, RWTH Aachen, Eilfschornsteinstraße 18, 52062 Aachen, Germany

Corresponding author:: anja.dannewitz@fkie.fraunhofer.de

“a wealth of information creates a poverty of attention” (Simon, 1971, p.40).

The rapid rise of emerging technologies like Artificial Intelligence (AI) poses the risk that individuals and societies cannot adapt fast enough to the corresponding cognitive requirements. This potential desynchronization between different parts of societies is described by Rosa (2005/2013) as a result of acceleration challenging the decision-making process. How can humans still make good decisions when time is short, and information is incomplete or overwhelming?

As a potential solution, behavioral interventions can guide humans in their decision-making process. Hereby, the concepts of Nudging (Thaler & Sunstein, 2008), Digital Nudging (Weinmann et al., 2016) and Boosting (Grüne-Yanoff & Hertwig, 2016; Hertwig & Grüne-Yanoff, 2017) seem to be promising. (Digital) Nudges are interventions that steer behavior by influencing and designing aspects of the (digital) decision environment (Thaler & Sunstein, 2008; Weinmann et al., 2016). As an extension, Weßel et al. (2019) proposed Self-determined Nudging in the human-machine interaction and cooperation to strengthen the freedom of choice by assessing the goals and aspirations of the individual and choosing the Nudges accordingly. Boosts aim to enhance decision-making competences and thereby empower people to make better choices by themselves (Hertwig & Grüne-

Yanoff, 2017). Or as Grüne-Yanoff et al. (2018) framed it, Nudges intervene on the individual's choice environment whereas Boosts intervene on their heuristic repertoires. These concepts have an explicit potential to increase resilience and adaptability to technologically enabled developments. The implementation must be guided by moral and ethical principles to maintain the autonomy of the human.

We argue that combining Cognitive Boosting and Nudging (CoBoNu) as a complementary approach can improve the intentional and rational decision-making process fundamentally.

Towards the CoBoNu framework

In Economics the homo oeconomicus describes the ideal human, who acts rational and maximizes utility in economic decisions (see for example Kirchgässner, 2008). However, human decision-making is not as rational as often assumed. The concept of Bounded rationality is one of the foundational principles in behavioral economics (Simon, 1955, 1990). Individuals are biased due to limited time, knowledge and capacities, and therefore use heuristics to overcome these limitations (Tversky & Kahneman, 1974). Kahneman (2011) differentiates therefore the thinking process into two systems: System 1 (fast, automatic, little or no effort) and System 2 (slow, conscious, effortful). Behavioral interventions like Boosts and Nudges facilitate individuals' decision-making but differ

in their approach: while Nudges usually target immediate decision-making, Boosts aim to provide the necessary competences that can be used in a specific decision-making situation (Hertwig & Grüne-Yanoff, 2017; Thaler & Sunstein, 2008). We assume that a systematic combination of Boosts and Nudges (CoBoNu) has great potential to enhance the individual decision-making process.

One common decision-making model in military research is the OODA (Observe – Orient – Decide – Act) loop (Boyd, 1996, 2018). Figure 1 shows how the CoBoNu approach could influence different phases of the OODA loop supported by a smart AI enabled tool. Based on the concept of Self-determined Nudging (Weßel et al., 2019), the human communicates his goals and preferences to an AI enabled tool which then develops and applies corresponding Boosts and Nudges. Boosts will especially address the Observe and Orient Phase, and Nudges will influence more greatly the Decide Phase. Additionally, the proposed approach will use Boosting as a pre-stage intervention fostering long-term habit formation by enhancing the individuals' decision-making capabilities. In a concrete choice situation, Boosts will be followed or accompanied by Nudges to support the individual. Moreover, it is supposed to be used iteratively to enhance the decision-making process.

Why do we need Boosts and Nudges?

The study of Igei et al. (2024) highlights the positive synergistic effect of both concepts in an energy saving education program, showing that the combination can create long-term behavior-change effects. Based on Ku et al. (2026), who demonstrated that Boosts and Nudges in the context of misinformation can promote critical thinking, we plan to explore how the CoBoNu approach can enhance the security-related decision-making process. A combination and/or complementary synergy of Boosts and Nudges may increase individuals' resilience. Thereby the advantages of both concepts will enable bet-

ter and informed judgements. One does not rule out the other; rather, they will be used in a way that builds on one another. Ono (2023) also developed the idea to combine Nudges and Boosts in personal AI agents. Hence, personalized adaptation, supported by AI agents, might be the key to addressing the needs and aspirations of the decision-maker. In the following, two examples are illustrated to demonstrate possible domains of use.

Alarming developments in the information domain are fake news, misleading information, or AI manipulated content that are spread especially via different media channels threatening society and democratic systems. Also, the potential danger of automation bias (Romeo & Conti, 2026) and validating machines (Krikorian, 2025) is increasing with more advanced technologies. Therefore, the question to be answered is how choice architects can design more confidence building measures to enhance skills like critical thinking and increase the user's competence to verify the accuracy of the information? As manipulated information is also an activity in the cognitive warfare domain, it is crucial to enable decision-makers to decide in the best way possible based on trustworthy and reliable information. Boosting and Nudging are possible interventions against misinformation on the individual level (Ku et al., 2026; Roozenbeek et al., 2023). Herzog & Hertwig (2025), based on McGrew (2024), also recommend boosting lateral reading to counteract misinformation. Hence, the CoBoNu framework can be applied, for example, to debunk misleading information by developing materials, which explain how to recognize false information and enhance critical thinking and lateral reading (Boosts). In a specific decision situation, individuals can use their acquired skills to not trust and share false content, in conjunction with another Boost that could provide quick instructions on how to assess the accuracy. To reinforce the competences, cues could pop up asking for critical examination of the provided content (Nudges). Thus, the combination of short-term intervention and ong-term

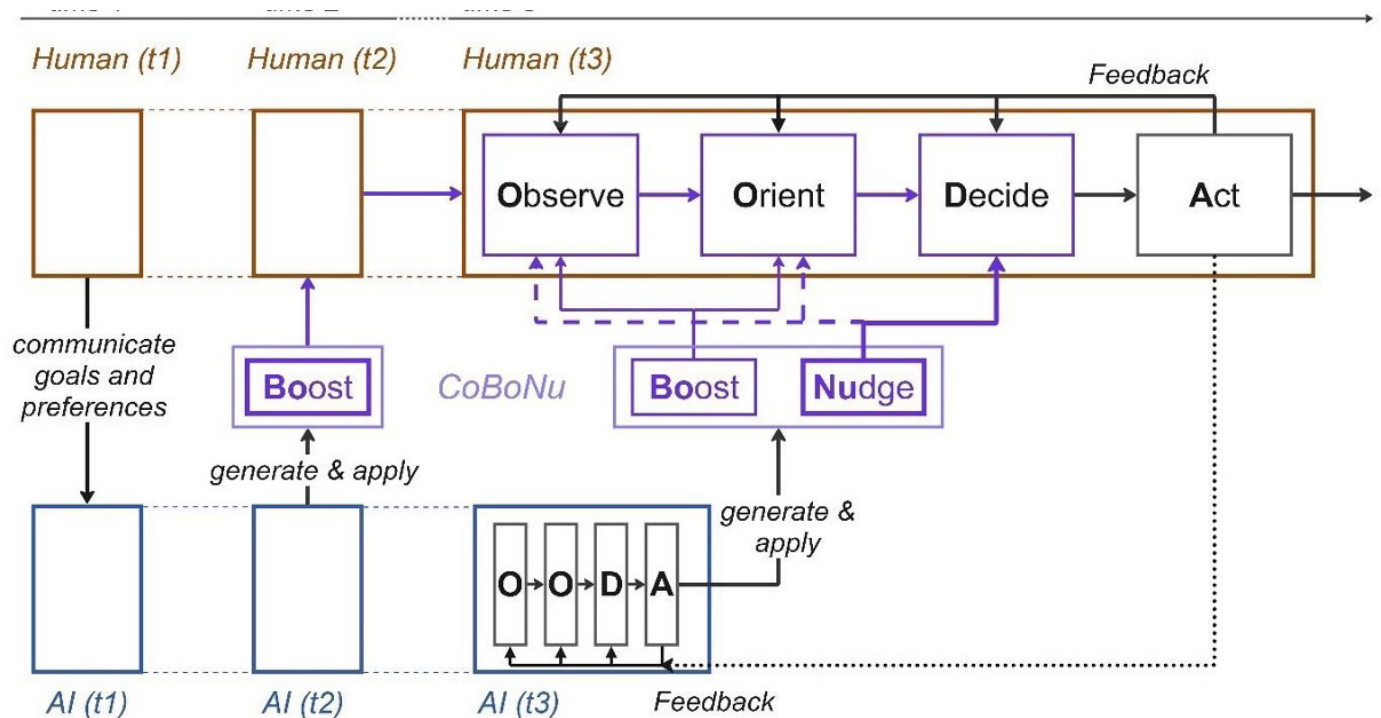


Figure 1: Visualization on how the CoBoNu approach could be integrated into the interaction between a human and an AI enabled tool, based on the OODA loop (Boyd, 1996, 2018) and nested OODA loops for decision-making (Flemisch et al., 2024)

capability development could provide users with the required attentive point of view.

As the battlefield is changing with new technologies, decision-making is also evolving, and operators need to adapt accordingly. Especially in military operations, decision-makers often face time pressure and unpredictability (e.g. D'Alessio et al., 2024). The use of digital systems might change the way decisions are made. Even more than ever, operators decide without any direct physical contact with the tangible consequences of the decision made. To ensure that humans remain in control, the concept of Meaningful Human Control (MHC) needs to be deliberately designed into these systems (Santoni de Sio & van den Hoven, 2018; Mecacci et al., 2024). The CoBoNu approach could facilitate and strengthen the implementation of MHC elements into these systems. Nudges might provide necessary reminders or cues to focus the operator's limited attention in a stressful decision moment. Based on the research of Becker et al.

(2022), automatically generated decision-aids by AI can be used to boost and improve human decision-making. Hence, such AI-generated boosts may also facilitate more responsible and effective usage of systems supporting the individual learning process.

In such discourses it is crucial to differentiate between requirements and domains to balance demand and support accordingly. Thereby, CoBoNu interventions and an applied tool should be designed in an ethical manner to provide support without restricting freedom of choice. Policies must ensure that no paternalistic or improper interests are pursued and that the implementation is aligned with legal regulations. Transparency, independence and MHC should remain core principles of the CoBoNu framework and a possible applied digital tool. Further exploration of which specific behavioral interventions provide the most effective and ethical support, especially in security-related decision-making, will enable more individualized intervention designs.

The way ahead

We advise translating the CoBoNu approach into an applied digital tool. An example based on the idea of Ono (2023) could be a digital personalized AI agent/interface that incorporates ethical standards and is developed as a complex system proposing the optimal interventions aligned with the user's goal. Notably, personal adaptation is crucial, because users perceive information differently.

Before conducting a quantitative study, interviews and an exploratory workshop should be organized to discuss and prototype such a personalized agent/interface using the Human Systems Exploration method (Flemisch et al., 2022; Preutenborbeck & Flemisch, 2026). The paradigm of Human Systems Exploration was inspired by the Exploratory Teams of the NATO Science and Technology Organization and allows different stakeholders to discuss necessary requirements and possible designs of such a proposed tool. Using this method provides an interdisciplinary technique to identify necessities and design ideas.

Concluding, the key advantage of the CoBoNu approach is to support individuals' decision-making in security-related situations by a combined approach of fostering their decision-making competences and nudging towards the self-determined goal. A better understanding of how Boosts and Nudges can be used in those situations will positively contribute to the Future of Decision-Making project. It can especially be used as an iterative guide to develop and enhance existing support tools. Also, it can increase awareness and strengthen resilience. Further implementation efforts and discussions on ethical eligibility are required.

Only with such ethical, value-based approaches as CoBoNu and its associated assistance systems, future decision-makers can balance human and technological characteristics and benefit from the technological progress of AI.

References

- Becker, F., Skirzyński, J., van Opheusden, B., & Lieder, F. (2022). Boosting human decision-making with AI-generated decision aids. *Computational Brain & Behavior*, 5, 467–490. <https://doi.org/10.1007/s42113-022-00149-y>
- Boyd, J. R. (1996, January). The Essence of Winning and Losing. Unpublished presentation. https://ooda.de/media/john_boyd_-_the_essence_of_winning_and_losing.pdf
- Boyd, J. R. (2018). A discourse on winning and losing (G. T. Hammond, Ed.). Air University Press. https://www.airuniversity.af.edu/Portals/10/AUPress/Books/B_0151_Boyd_discourse_Winning_Losing.pdf
- D'Alessio, I., Aitella, U., Giannini, A. M., & Burrai, J. (2024). "What about military decision-making?": A bibliometric review of published articles. *Behavioral Sciences*, 14(7), 514. <https://doi.org/10.3390/bs14070514>
- Flemisch, F., Preutenborbeck, M., Baltzer, M., Wasser, J., Kehl, C., Grünwald, R., Pastuszka, H.-M., & Dahlmann, A. (2022). Human systems exploration for ideation and innovation in potentially disruptive defense and security systems. In G. Adlakha-Hutcheon & A. Masys (Eds.), *Disruption, ideation and innovation for defence and security* (pp.79–117). Springer. https://doi.org/10.1007/978-3-031-06636-8_5
- Flemisch, F., Preutenborbeck, M., Burlage, L., & Ripkens, A. (2024). Dilemma model of cooperative guidance and control to solve the speed-certainty-dilemma in human-autonomy teaming: First sketch. 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS), 1–4. <https://doi.org/10.1109/ICHMS59971.2024.10555733>
- Grüne-Yanoff, T., & Hertwig, R. (2016). Nudge versus boost: How coherent are policy and theory? *Minds and Machines: Journal for Artificial Intelligence, Philosophy and*

Cognitive Science, 26(1-2), 149–183. <https://doi.org/10.1007/s11023-015-9367-9>

Grüne-Yanoff, T., Marchionni, C., & Feufel, M. A. (2018). Toward a framework for selecting behavioural policies: How to choose between boosts and nudges. *Economics and Philosophy*, 34(2), 243–266. <https://doi.org/10.1017/S0266267118000032>

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, 12(6), 973–986. <https://doi.org/10.1177/1745691617702496>
Herzog, S. M., & Hertwig, R. (2025). Boosting: Empowering citizens with behavioral science. *Annual Review of Psychology*, 76, 851–881. <https://doi.org/10.1146/annurev-psych-020924-124753>

Igei, K., Kurokawa, H., Iseki, M., Kitsuki, A., Kurita, K., Managi, S., Nakamuro, M., & Sakano, A. (2024). Synergistic effects of nudges and boosts in environmental education: Evidence from a field experiment. *Ecological Economics*, 224, 108279. <https://doi.org/10.1016/j.ecolecon.2024.108279>

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
Kirchgässner, G. (2008). *Homo oeconomicus: The economic model of behaviour and its applications in economics and other social sciences*. Springer.

Krikorian, R. (2025, October 30) *The Validation Machines*. The Atlantic. <https://www.theatlantic.com/ideas/archive/2025/10/validation-ai-raffi-krikorian/684764/>

Ku, K. Y. L., Li, J., Lun, V. M. C., Tsang, Y. K., Lui, M., Kong, Q., & Song, Y. (2026). Empowering critical thinking in combating social media misinformation: Testing efficacy of digital “boost and nudge” interventions. *Computers in Human Behavior*, 175, 108829. <https://doi.org/10.1016/j.chb.2025.108829>

Mecacci, G., Amoroso, D., Cavalcante

Siebert, L., Abbink, D., van den Hoven, J., & Santoni de Sio, F. (2024). *Research handbook on meaningful human control of artificial intelligence systems*. Edward Elgar Publishing.

McGrew, S. (2024). Teaching lateral reading: Interventions to help people read like fact checkers. *Current Opinion in Psychology*, 55, 101737. <https://doi.org/10.1016/j.copsyc.2023.101737>

Ono, T. (2023). Nudge & boost agents: Designing ambient intelligent systems to effectively influence human decision making. 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2094–2099. <https://doi.org/10.1109/RO-MAN57019.2023.10309371>

Preutenborbeck, M., & Flemisch, F. (2026). Human-systems dialectics and exploration: Balancing system design thinking for the agile and participatory integration of sociotechnical systems. In G.A. Boy (Ed.), *Handbook of sociotechnical systems: A human systems integration approach* (2nd ed.). CRC Press.

Romeo, G., & Conti, D. (2026). Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. *AI & Society*, 41, 259–278. <https://doi.org/10.1007/s00146-025-02422-7>

Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering misinformation: Evidence, knowledge gaps, and implications of current interventions. *European Psychologist*, 28(3), 189–205. <https://doi.org/10.1027/1016-9040/a000492>

Rosa, H. (2013). *Social acceleration: A new theory of modernity*. (J. Trejo-Mathys, Trans.). Columbia University Press. (Original work published 2005)

Santoni de Sio, F., & van den Hoven, J. (2018). *Meaningful human control over autonomous*

systems: A philosophical account. *Frontiers in Robotics and AI*, 5, Artikel 15. <https://doi.org/10.3389/frobt.2018.00015>

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118. <https://doi.org/10.2307/1884852>

Simon, H. A. (1971). Designing organizations for an information-rich world. In M. Greenberger (Ed.), *Computers, Communications, and the Public Interest* (pp.37–52). The Johns Hopkins Press.

Simon, H. A. (1990). Bounded rationality. In: J. Eatwell, M. Milgate, and P. Newman (Eds.), *Utility and Probability* (pp.15–18). Palgrave Macmillan. https://doi.org/10.1007/978-1-349-20568-4_5

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185 (4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>

Weinmann, M., Schneider, C., & vom Brocke, J. (2016). Digital nudging. *Business & Information Systems Engineering*, 58(6), 433–436. <https://doi.org/10.1007/s12599-016-0453-1>

Weßel, G., Altendorf, E., Schwalm, M., Canpolat Y., Burghardt, C., & Flemisch, F. (2019). Self-determined nudging: a system concept for human–machine interaction. *Cognition, Technology & Work*, 21, 621–630. <https://doi.org/10.1007/s10111-019-00556-5>

Presenters



Daniel Lucuix studies International Relations at Stanford University, where he is also pursuing an M.A. in Public Policy with a concentration in AI governance and national security. This year he was selected as a mentee for the Hoover Institution's National Security Affairs Fellows mentorship program. His research has ranged from global AI legislation, which he mapped with his peers for Stanford's International Relations Policy Lab, to nineteenth-century French colonial slave registers; the latter led to a co-authored article now under review at *The Journal of African History*. He grew up in Tampa, Florida, speaks French at a professional level, and has a passion for public service.



Eli Talbert, Ph.D., is a Lead AI and Decision Engineer at MITRE, where he works on the evaluation, adoption, and deployment of artificial intelligence in high-stakes government settings. Previously, he served as a senior data scientist supporting U.S. Special Operations Command and conducted research at the RAND Corporation. A former National Science Foundation Graduate Research Fellow, his work spans applied machine learning, AI governance, and institutional decision-making. His writing has appeared in *War on the Rocks*, *The Regulatory Review*, and *Tech Policy Press*.



Dr. Rachel A. George is a Lecturer in International Relations at Stanford University and a Legal Innovation & Policy Fellow at Stanford University, where her research focuses on how institutions adapt to political and technological change. She is also an affiliate with Stanford's Center for International Security and Cooperation and Nonresident Scholar with the Carnegie Endowment for International Peace. Previously, she served as Director of Education Content at the Council on Foreign Relations, a Lecturing Fellow at Duke University's Sanford School of Public Policy, and a Fellow in the Georgetown Institute for Women, Peace and Security. She holds a PhD in International Relations from the London School of Economics and Political Science, an MA in Middle Eastern Studies from Harvard University, and a BA in Politics from Princeton University.



Dr Tim Jones is a coach developer, educator and researcher with over 20 years of experience across professional and grassroots sport in the UK, US and Middle East. His work focuses on sport coach development and the relationship between learning, reflection and practice.

Holding a PhD in Coach Development, Tim brings together academic research and practical experience to explore how individuals learn, navigate complexity and make decisions in real-world environments. His particular interests include critical realism, adult development, and the evolving role of artificial intelligence in supporting sport coaching professional judgement.

Presenters



Andrew Bell is a Nonresident Fellow with the Military Ethics Research Lab and Innovation Network (MERLIN) at UNSW Canberra and an Affiliate of Stanford University's Center for International Security and Cooperation (CISAC). His interdisciplinary research sits at the intersection of international security, international law and ethics. He studies how norms and the law of armed conflict shape military conduct and how commanders make decisions about the use of force. His current work examines how AI decision support systems are reshaping command decision-making and the allocation of risk in war. His work has appeared in *Perspectives on Politics*, *Security Studies*, the *Journal of Peace Research*, as well as in the *Washington Post*, *Newsweek*, *The Hill*, and *Foreign Policy*, among other outlets.

Most recently, Dr. Bell was a senior research fellow at Stanford's CISAC. Before that, he served as program lead for Professional Military Education at the U.S. Department of Defense's Civilian Protection Center of Excellence. He has also been a senior research fellow at the International Committee of the Red Cross, a visiting professor at the U.S. Army War College, and an assistant professor at Indiana University's Hamilton Lugar School of Global and International Affairs. He has held fellowships at the U.S. Military Academy, the U.S. Naval Academy, George Washington University's Elliott School of International Affairs, and the Centre for U.S. Politics at University College London. He is a retired U.S. Air Force lieutenant colonel with deployments to Iraq and Afghanistan. He holds a Ph.D. in political science from Duke University, a J.D.-M.A. from the University of Virginia School of Law, and an M.T.S. from Duke Divinity School.



Anja Dannewitz studied Business Administration at the University of Cologne, where she developed a strong interest in Psychology as a field of study. She completed her Master's degree in Psychology in Business and Economics at Universidade Católica Portuguesa in Lisbon. Her scientific focus lies in Behavioural Science, especially Decision-Making and Human Systems Exploration of cooperative human-machine systems. She is currently working as a research scientist at Fraunhofer FKIE, where she supports the Principal Scientist of Human Systems (Prof. Frank Flemisch).

Presenters



Indrek Sarap is an active duty army officer of the Estonian Defence Forces with 25 years of experience in various command and staff positions, including deployments to Afghanistan in 2006 and Mali in 2013-2014.

Indrek is an honour graduate of the Baltic Defence College and he has a masters degree from the Estonian Military Academy, where he measured the quality on military commanders solving tactical problems and utilizing two separate decision-making processes. Indrek continues his research in military decision-making as a visiting scholar at CISAC and as part of the Future of the Decision-Making project.



Paul Larcey co-directs the Princeton University Global Systemic Risk Research Group, where he researches large-scale risks in highly coupled systems, emerging technologies, and global supply chains, working alongside US and European policy and defence institutions. He also serves as a UK government Sector Specialist, shaping Innovate UK/UKRI's national security and deep tech strategy from both a technology and funding perspective and has a career spanning venture capital and private equity, including senior leadership roles in PE-backed firms.

Stanford

Geopolitics, Technology,
and Governance
Center for International Security and Cooperation



THE FUTURE OF DECISION_MAKING CONFERENCE
28th October 2026