Received: 27 February 2025 | Revised: 9 December 2025 | Accepted: 25 March 2026 | Published online: 28 April 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–13
DOI: 10.47852/bonviewAIA62025549

BON VIEW PUBLISHING

Exploration of Trust and Decision-Making in AI-Augmented Military Domain: A Framework for Human–Machine Teaming

Janar Pekarev[1,*] and Priit Värno[2]

[1] *Force Transformation Command, Estonian Defence Forces, Estonia*

[2] *Department of Applied Research, Estonian Military Academy, Estonia*

Abstract: This study presents a theoretical and methodological framework for examining trust and decision-making in AI-augmented human–machine teaming within military contexts. The framework combines a supervised learning backbone (a pre-trained Random Forest classifier) with a deterministic rule-based override layer that encodes non-negotiable constraints aligned with key laws and principles of armed conflict, including distinction, proportionality, and military necessity, as well as related rules-of-engagement logic. Using a structured generator of combat-relevant targeting scenarios, the system produces recommendations that can be accepted, escalated, vetoed, or deferred, allowing shifts in cognitive authority and reliance to be observed and measured. An interactive, scenario-driven interface exposes calibrated confidence, salient feature cues, and explicit override traces to support verification and controlled reliance in uncertain situations. On a large synthetic scenario corpus, the model exhibits stable performance and well-calibrated probability estimates. At the same time, the guardrail layer systematically redirects borderline engage outputs toward safer outcomes and audit-ready escalation states. The artifact is positioned as a research instrument rather than an operational decision-making authority. It is designed to elicit and quantify trust calibration behaviors, including cautious skepticism, confident alignment, and deliberative hesitation, across varying levels of complexity and ambiguity. The design is released for replication and iterative refinement, supporting interdisciplinary evaluation of transparent, doctrine-compatible AI decision support and providing a practical basis for controlled user studies on trust, bias, and ethical judgment in military human–AI teaming.

Keywords: AI-augmented judgment, trust calibration, human–machine teaming, military decision-support systems

1. Introduction

Machine learning is increasingly embedded in high-stakes decision-making. In medicine, AI-assisted diagnosis and treatment promise faster and more accurate assessments [1, 2], but adoption depends on perceived accuracy, explainability, and accountability [3]. Studies that manipulate clinical decision-support rationales (e.g., deliberately creating false positives/negatives) demonstrate the fragility of trust: confidence increases when justifications align with clinicians' knowledge and collapses when errors are exposed [4–6]. Whereas human error often erodes trust gradually, even minor algorithmic missteps can trigger abrupt distrust, underscoring the need to calibrate reliance on a case-by-case basis rather than grant blanket trust [7].

The military operates under similarly acute time pressure, ambiguity, and asymmetric risk and already employs AI for object detection, cybersecurity, robotics, logistics, and battle management [8]. Errors can range from resource misallocation to lethal consequences [9, 10], and overreliance can hide critical failures, such as false negatives in cyber defense, where operators assume the system is correct despite a compromise [11]. Properly designed systems can reduce cognitive load and highlight weak signals, thereby improving collaborative decision-making and risk assessment [12]. Just as CT/MRI provide anchors for evaluating clinical AI, military operations offer quantifiable proxies, such as real-time imagery, battle damage assessment, and rules of engagement (ROE), which can structure evaluation [13]. AI-enabled battlefield management is also shifting the commander's role and proximity to the fight, challenging purely intuition-driven decision-making and increasing demands for auditable rationales [13].

Cross-domain evidence identifies three levers for calibrated reliance: (i) surface uncertainty rather than hide it; (ii) provide concise, verifiable rationales; and (iii) enforce non-negotiable constraints as rules, not soft preferences [14, 15]. Overreliance, however, can produce automation bias. In one study, agreement with binary AI diagnoses dropped from 86.7% to 45.8% when heatmap rationales were shown, indicating miscalibration rather than repair [16]. For military teaming, the implication is not to reject explanations but to present them in forms that support verification and control, maintaining human agency while avoiding uncritical acceptance [17–19].

***Corresponding author:** Janar Pekarev, Force Transformation Command, Estonian Defence Forces, Estonia. Email: janar.pekarev@mil.ee

Trust in human–machine teams is multifaceted and socio-technical, comprising cognitive, behavioral, physiological, and computational elements that jointly shape reliance [20–23]. Predispositions, prior experience, and perceived reliability influence expectations and behavior [24]. In adjacent clinical contexts, operators adapt their reliance as they encounter error patterns or opaque judgments [25, 26]. Robust evaluation, therefore, requires multiple lenses: subjective measures (self-reports, interviews, focus groups) and objective measures (behavioral indicators in simulations, reliance/hesitation/error tolerance under time pressure) [27–29]. Computational and system-level metrics then link operator behavior back to model performance and explainability quality [21].

Despite progress in other domains, the military context lacks a dedicated, research-purpose platform for studying AI-augmented decision-making at the level of doctrinal constraints and operator behavior. Recent work using large language models to simulate decisions offers convenience but typically lacks domain-specific military training, which risks oversimplifying the dynamics of combat decisions [30]. Until such models are augmented with targeted military data and embedded in explicit evaluative scaffolding, their utility remains mostly theoretical and prone to overclaim [30].

To address this gap, we present a policy-aware teaming framework and an openly available toolchain that enable structured interaction with AI recommendations in randomized targeting scenarios. The framework couples a supervised learner with a rule-based override layer that encodes protected-asset constraints, civilian-harm thresholds, strategic-infrastructure sensitivities, and escalation controls, all of which are aligned with doctrine and international humanitarian law (IHL). The operator interface presents final recommendations (Do Not Engage, Do Not Know, Ask Authorization, Engage), calibrated confidence levels, concise feature-based rationales, and explicit flags when overrides are activated. The tool is positioned as a research instrument rather than a command system; its purpose is to measure and shape human judgment, not to replace it [20–23].

Methodologically, it supports (i) scenario simulation to probe decision boundaries, (ii) performance measurement that links operator reliance to model reliability, and (iii) override detection to capture where expert judgment contests algorithmic advice. In targeting, where "selecting and prioritizing targets and matching appropriate responses" is governed by formal doctrine, we deliberately maintain visibility and auditability of doctrinal alignment [31]. The implementation is publicly released to support replication, structured critique, and ethically bounded experimentation, consistent with broader aims for transparent and explainable AI [32].

Concretely, this paper (a) formalizes a policy-aware pipeline that pairs a supervised learner with a rule-based override layer aligned to doctrine and IHL; (b) constructs a 36 feature scenario space to stress test recommendations under shifts in observability, ambiguity, and political sensitivity; (c) standardizes reporting to foreground per class metrics, reliability, and how guardrails reshape safety-critical errors; and (d) specifies a user-study plan for trust calibration under cognitive load. We restrict cross-domain content to this brief bridge; all empirical focus and evaluation criteria are military-specific, drawing on established work on trust, explainability, and human–AI interaction [14–22, 24–29]. Section 2 frames trust, agency, and guardrails. Section 3 details scenario generation, models, overrides, interface, and evaluation protocol. Section 4 reports results, including baselines, rule trigger frequencies, sensitivity analyses, and planned user validation. Section 5 concludes with implications for doctrine-compatible AI and future work.

2. Theoretical Framework

Military decision support is judged under three non-negotiable criteria: time pressure, accountability to doctrine and law, and uncertainty that is rarely stationary. In this environment, "trust in AI" is not a feeling but a control policy that allocates cognitive authority between the human operator and automation as operational/data complexity changes. This framework aims to make that policy explicit and testable. It does so by defining a decision surface, enumerating discrete decision states, and specifying guardrails that bind automation to ROE and humanitarian law (IHL).

2.1. The trust calibration surface

Figure 1 introduces the calibration surface on the trust domain. The horizontal axis encodes operational/data complexity (quality, volume, and conflict of evidence; Electronical warfare (EW) interference; collateral-risk factors). The vertical axis represents cognitive authority, ranging from AI autonomy to shared control and human oversight. Figure 1 illustrates the calibrated ridge and the four decision states connected along a single trajectory. The dashed trajectory shows the intended progression of decision states as conditions change: Engage → Ask Authorization → Do Not Engage → Do Not Know.

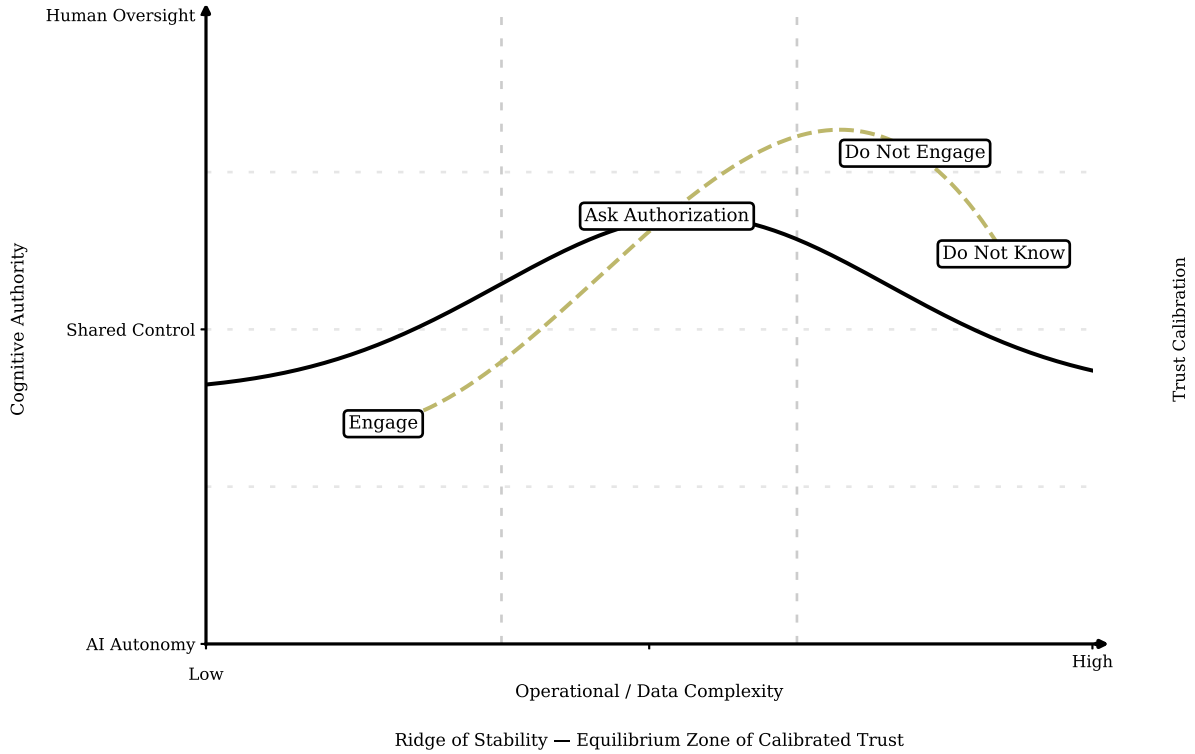
The solid ridge is the zone of calibrated trust, a combination of complexity and authority, where the human–AI team performs best. The dashed trajectory is not decorative; it is the decision continuum the operator is expected to traverse as conditions change. It connects four labeled decision states used throughout the paper and implemented in the system:

1) Engage: conditions are low-complexity and well-observed; model confidence is high; ROE/IHL constraints are not triggered. Authority permissibly tilts toward automation.
2) Ask Authorization: complexity and ambiguity rise (e.g., proximity to protected entities, incomplete Positive Identification (PID), contested sensors). Authority is shared; escalation to a human authority is required by doctrine.
3) Do Not Engage: high complexity and/or risk: conflicting evidence, protected objects, or uncertain PID. Human oversight dominates; ROE/IHL guardrails veto fire.
4) Do Not Know: novelty or out-of-distribution (OOD) conditions that neither model confidence nor doctrine can resolve at the current time scale. The correct action is to withhold commitment and seek clarification or additional Intelligence, Surveillance, and Reconnaissance (ISR).

Along this trajectory, Figure 1 represents a policy for trust migration: as complexity increases, cognitive authority is expected to shift upward from the AI to the human, and reliance transitions from automated execution to supervision, escalation, or abstention. The framework comprises three interacting elements as follows:

1) AI decision-support module. A supervised classifier provides an initial recommendation and calibrated probability, complemented by an explanation interface that reveals salient factors. As argued in Section 3, this affords partial transparency

Figure 1
Trust calibration surface for human–AI teaming



sufficient to support operator judgment without claiming full interpretability.

2) Rule-based guardrails (ROE/IHL). Deterministic checks evaluate the context for hard constraints (protected categories, authorization requirements, out-of-domain combinations). These guardrails can override or gate the model output, making the automation accountable to doctrine.
3) Human decision maker. The operator retains final authority, exercises escalation paths, and provides feedback signals that are used to adjust thresholds and presentation (e.g., more rationale under high ambiguity). Human oversight is not a failsafe at the end of a pipeline; it is an active control layer whose allocation is visualized in Figure 1.

2.2. Decision states and trust management

The four decision states are not free-floating labels. They are anchored to measurable, auditable, and reproducible triggers. Engage → Ask Authorization is triggered when uncertainty rises above a calibrated band, when sensitive context appears (e.g., dual-use infrastructure, proximity to civilians), or when the model's reliability estimate degrades. The system surfaces rationale and flags the requirement for human confirmation. Ask Authorization → Do Not Engage is triggered when any hard guardrail activates (protected object, violated PID standard, unacceptable collateral risk) or when human rationale contradicts the automation with sufficient justification. Any state → Do Not Know is triggered by OOD evidence, incoherent sensor fusion, or adversarial conditions (e.g., spoofed tracks).

The correct behavior is to express explicit uncertainty, defer, and seek information rather than remain silent and fail. These transitions are the behavioral semantics of the dashed trajectory in Figure 1. They also define what we measure later: override frequencies, escalation rates, and how often authority migrates as designed rather than sticking (automation bias) or collapsing (over-caution).

Operationally, each decision cycle proceeds as follows: the AI proposes a class and confidence; guardrails check doctrine; the User Interface (UI) reveals both recommendation and any guardrail trace; the operator either agrees, escalates, vetoes, or declares uncertainty; the outcome is logged. Two feedback channels are for update reasons. At higher complexity, the system increases transparency (by highlighting salient features, presenting alternative options, and indicating uncertainty ranges), while at lower clarity, it compresses into concise cues to preserve tempo. Threshold tuning is set to aggregate evidence from overrides and errors, repositioning the calibrated band that defines the ridge and shifting the team toward earlier escalation in ambiguous sectors and earlier engagement in well-understood sectors. The loop is designed to prevent both misuse and disuse: neither blind acceptance of automation nor habitual rejection should be a stable equilibrium.

2.3. Testable propositions

The framework yields concrete propositions for empirical work:

1) Calibration migration. As operational/data complexity increases, the probability of "Ask Authorization" and "Do Not Engage" rises, while "Engage" declines; "Do Not Know" concentrates at the far right tail.
2) Guardrail efficacy. Hard-constraint rules reduce critical false positives at a small cost to the engagement rate; vetoes cluster in contexts that the doctrine intends to protect.

3) Transparency under load. Rationale improves alignment with the calibrated band under moderate complexity but has diminishing returns or even harms tempo under low complexity.
4) Operator learning. Over repeated cycles, escalation and veto behavior converge toward the dashed trajectory, reducing both automation bias (over-trust) and complacent abstention (under-trust).

These propositions directly motivate the user-study design outlined later (rationale shown vs hidden; low vs high cognitive load; outcomes: agreement rate, latency, and a trust-calibration index). The trust calibration surface framework does not imply that one curve fits all missions. The ridge's shape is mission- and ROE-dependent; Figure 1 is a policy template, not a statistical fit, nor is the classifier "interpretable" by itself; transparency is achieved through design choices (guardrails, rationale exposure, standardized reporting) rather than by model mysticism.

3. Research Methodology

3.1. System architecture and scenario generation

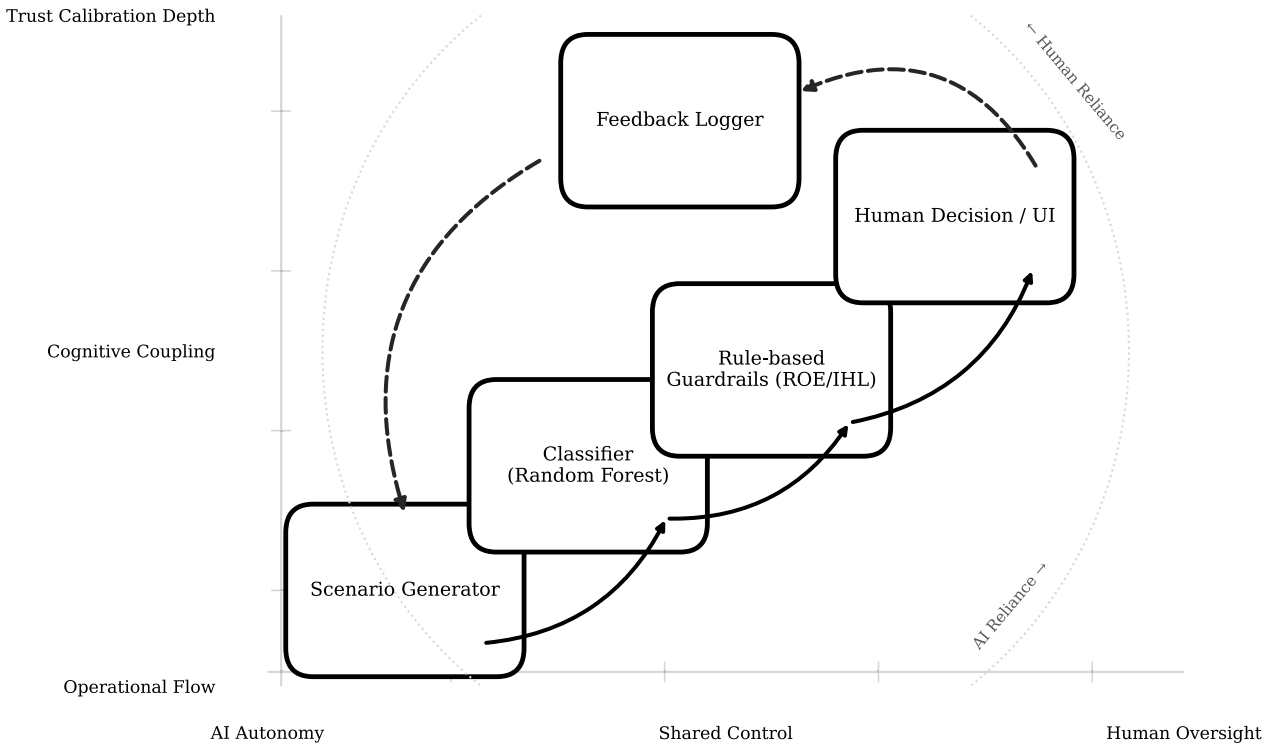
The study utilizes a structured generator to produce realistic and auditable targeting scenarios for experimentation. Each scenario is represented as a feature vector (36 fields) encompassing target characteristics, terrain, civilian presence, weaponeering, proximity to protected objects, collateral damage potential, and flags indicating political sensitivity, legal advice, and ethical concerns. Scores for distinction, proportionality, and military necessity are included from both an AI "assessment" and a human-notional assessment to induce convergent and divergent cases. Randomization operates over coupled feature groups rather than independent sampling to preserve face validity (e.g., a medical installation implies protected-object flags and high collateral risk). The generator outputs the feature set, along with a Total_Score heuristic that summarizes the risk-benefit for presentation to the operator and for rule triggers in the override layer.

Figure 2 illustrates the control-loop architecture and the interactions among these components. In Figure 2, modules are positioned along two conceptual axes: the x-axis represents the balance between automation and human oversight. In contrast, the y-axis represents the increasing depth of trust calibration. The feedback logger closes the loop by feeding operator responses and override data back to the scenario generator, completing the trust-calibration cycle shown in Figure 2. The generator spans the operational/data-complexity axis in Figure 1 by varying evidence quality, context conflict, and collateral constraints, ensuring coverage across low-ambiguity to high-ambiguity regimes, which are later used to test trust migration along the dashed trajectory.

3.2. Data preprocessing and model training

The corpus comprises 1 million synthetic scenarios with labels reflecting the decision a competent commander should take under ROE/IHL. The learned classifier predicts a binary action (Engage vs Do Not Engage). Intermediate actions (Ask Authorization; Do Not Know) are generated by the rule-guarded layer, not by the learner, to maintain doctrinal authority explicitly. Categorical fields are one-hot encoded; continuous scores are scaled to a standard numeric range. The dataset is split 80/20 for

Figure 2
Control loop architecture for AI-augmented decision support



Solid arrows: operational and decision flow. Dashed arrows: trust calibration feedback loop.
Modules are positioned by automation–oversight balance (x) and trust-depth hierarchy (y).

training/testing with class balance maintained across splits. A Random Forest (scikit-learn) is trained with 100 trees, the Gini criterion, a depth cap (max_depth = 20), and default regularizers. Training on 800k scenarios completes within desktop CPU budgets; evaluation uses the 200k hold-out set plus 5-fold Cross-Validation (CV) on the training partition. Headline results (full details in Section 4): 95.86% accuracy on the held-out set; cross-validation accuracy 95.77% mean with low variance. Class-wise precision/recall/F1 and F1-scores, along with a confusion matrix, are reported in Section 4.1, accompanied by a reliability snapshot.

3.3. Override rule module and selection rationale

Rules implement ROE/IHL and the commander's intent as prioritized checks:

Rule 1 – Do Not Engage. Protected objects, unlawful targets, or extreme civilian-risk thresholds trigger an immediate veto, independent of model confidence.

Rule 2 – Ask Authorization. Ambiguous or politically sensitive contexts require escalation (e.g., dual-use infrastructure, uncertain PID, contested sensors).

Rule 3 – Do Not Know. OOD or incoherent combinations (e.g., weapon–target mismatches, spoofed tracks) require explicit uncertainty and information-seeking.

Rule 4 – Engage. If no rules fire, the model's recommendation stands.

Each rule is a documented Boolean condition evaluated in priority order; the system logs which rule fired for auditability and later analysis (Section 4.2 reports trigger frequencies and effects). This design reflects the "ethical governor" concept as a deliberate safety bottleneck. Random Forests perform strongly on structured tabular data, capture nonlinear interactions, train efficiently, and support partial transparency via global importance and post hoc local explanations. They are more analyzable than deep nets for this task, without claiming inherent instance-level interpretability. Baselines, such as logistic regression and Gradient-Boosted Trees (for instance, XGBoost), and a shallow Multilayer Perceptron (MLP) with SHAP/LIME are trained on identical splits. Their results are reported in Section 4.1 to justify the final choice in terms of performance, transparency burden, and computational/time.

3.4. Experimental procedure and interactive apps

No human participants were involved in the present study. No participant recruitment, informed consent, or human-subject data collection took place for the present study; all results reported here are based on synthetic scenarios and system-level evaluation only. The following describes a planned future experimental design for validation purposes. In a planned study, participants receive brief guidance and then complete up to 10 scenarios per session. In the first phase, participants, without seeing the model's rationale, would decide to establish a human baseline; in the second phase, the UI reveals the model's prediction, salient factors, and any rule overrides. The order can be counterbalanced. Each scenario has a fixed response window (e.g., 300 s) with auto-submit safety. The system would log action choice, latency, agreement with the model, escalation/veto events, and rule triggers. These metrics directly operationalize transitions along the dashed trajectory in Figure 1.

The main experiment app serves randomized scenarios; a second interactive tool allows manual scenario composition to probe sensitivity and gather qualitative rationales. Collected traces underpin the trust-calibration analyses and the user-study plan in Section 4. Table 1 specifies the participant-trace schema captured by the UI (decision, model prediction, latency, agreement, and free-text rationale), which underpins the trust-calibration analyses planned in Section 4.6. A public, synthetic-only demo of this decision-support interface is available at https://military.streamlit.app/ and https://human-machine-teaming.streamlit.app/, allowing readers to inspect the layout, scenario fields, recommendations, and guardrail indications directly.

To make the mechanics concrete, a single randomly generated scenario trace captured in the research app is summarized in Table 2. The trace links structured features through the model's output to any applicable guardrail activation and the final decision.

Complementing this internal record, Table 3 presents the corresponding user-interface rationale for the same scenario,

Table 1
Research data and feedback

Scenario	Participant decision	Model prediction	Decision time (s)	Confirmation feedback	Additional feedback
Target_Category: Bridging Unit, Score: 3 Civilian_Presence: 1–10, Score: -2 Weaponeering: 120 mm Mortar, Score: 3 Ethical_Concerns: Yes, Score: -2 AI_Distinction (%): 34, Score: -2 Human_Distinction (%): 90, Score: 4 Total Score: 17	Engage	Ask Authorization	122.9	Neither Agree Nor Disagree	"I would recommend allowing additional time to consider all aspects thoroughly..."

Note: Tables 1–3 present illustrative synthetic/interface examples for the planned study workflow; they are not data collected from human participants.

Table 2
Research data and feedback (randomly generated scenario)

Scenario	Parameters	Scores	Parameters	Scores
Target_Category:	Unmanned ground vehicle	1 (2.22%)	Command post	4 (7.27%)
Target_Vulnerability:	Low	2 (4.44%)	Very_High	5 (9.09%)
Terrain_Type:	Village	–3 (–6.67%)	Medical Facility	–5 (–9.09%)
Civilian_Presence:	30–49	–3 (–6.67%)	50–99	–4 (–7.27%)
Damage_Assessment:	Low	1 (2.22%)	Low	1 (1.82%)
Time_Sensitivity:	High	3 (6.67%)	Immediate	5 (9.09%)
Weaponeering:	Surface-to-Air missile	5 (11.11%)	Laser-Guided-Rocket	5 (9.09%)
Friendly_Fire:	Medium	1 (2.22%)	Very_Low	3 (5.45%)
Politically_Sensitive:	Very_Low	2 (4.44%)	Very_High	–1 (–1.82%)
Legal_Advice:	Questionable	1 (2.22%)	It depends	1 (1.82%)
Ethical_Concerns:	No	3 (6.67%)	Potential	1 (1.82%)
Collateral_Damage_Potential:	Medium	1 (2.22%)	Very_High	–3 (–5.45%)
AI_Distinction (%):	92	5 (11.11%)	83	4 (7.27%)
AI_Proportionality (%):	34	–2 (–4.44%)	57	1 (1.82%)
AI_Military_Necessity:	Yes	2 (4.44%)	Open to Debate	1 (1.83%)
Human_Distinction (%):	75	5 (11.11%)	100	5 (9.09%)
Human_Proportionality (%):	100	3 (6.69%)	50	–4 (–7.27%)
Human_Military_Necessity:	Yes	2 (4.44%)	Open to Debate	2 (3.64%)
Total score:	Scenario nr 1. Shuffled app	33	Scenario nr 2. User-interface app	21
Model prediction:	Engage		Do Not Engage *Note: High risk with a civilian presence*.	

Table 3
Research data and feedback (user interface)

Scenario	Parameters	Scores	Parameters	Scores
Target_Category:	Fighter Aircraft	4 (11.43%)	Electronic Warfare Installation	3 (5.88%)
Target_Vulnerability:	Low	2 (5.71%)	High	4 (7.84%)
Terrain_Type:	Electric Power Grid Network	1 (2.86%)	Checkpoint	1 (1.96%)
Civilian_Presence:	0	0 (0.00%)	1–10	–1 (–1.96%)
Damage_Assessment:	Low	1 (2.86%)	Very_High	3 (5.88%)
Time_Sensitivity:	Normal	1 (2.86%)	Normal	1 (1.96%)
Weaponeering:	155mm Artillery	4 (11.43%)	Cyber Weapon	5 (9.8%)
Friendly_Fire:	Medium	1 (2.86%)	Very_Low	3 (5.88%)
Politically_Sensitive:	Low	1 (2.86%)	Very_Low	2 (3.92%)
Legal_Advice:	Lawful	3 (8.57%)	Neutral	2 (3.92%)
Ethical_Concerns:	Plausible	1 (2.86%)	Hypothetical	1 (1.96%)
Collateral_Damage_Potential:	Low	1 (2.86%)	Medium	1 (2.0%)
AI_Distinction (%):	42	–1 (–2.86%)	100	5 (9.8%)
AI_Proportionality (%):	20	–4 (–11.43%)	100	5 (9.8%)
AI_Military_Necessity:	Yes	2 (5.71%)	Yes	2 (3.92%)

(*Continued*)

Table 3
(*Continued*)

Scenario	Parameters	Scores	Parameters	Scores
Human_Distinction (%):	70	1 (2.86%)	30	–5 (–9.8%)
Human_Proportionality (%):	50	–4 (–11.43%)	30	–5 (–9.8%)
Human_Military_Necessity:	Yes	3 (8.57%)	Open to Debate	2 (3.92%)
Total score:		17		29
Model prediction:	Scenario nr 3. Shuffled app	Do Not Know	Scenario nr 4. User-interface app	Ask Authorization

showing explanatory cues and override indicators as they appear to the operator. Together, these two sub-tables connect the system's internal logic with its human-facing explanation, illustrating how traceability and explainability jointly support later trust calibration analysis.

4. Evaluation and Discussion

This section characterizes the instrument on a synthetic corpus to demonstrate three key points: the predictive backbone is stable and well-calibrated; rule-based guardrails reshape errors in the intended safety-first direction; and the framework exposes measurable signals of trust behavior for later user studies. The results support the tool's validity as a measurement device, rather than an operational performance claim.

4.1. Quantitative performance on the synthetic corpus

All metrics are computed on the 200,000-scenario held-out test set described in Section 3.2. The classifier achieves 95.86% accuracy with balanced class performance. A 5-fold cross-validation on the training partition yields 95.77% (mean) with low variance, indicating that the predictive backbone is stable rather than tuned to a narrow slice of the synthetic distribution. The configuration from Section 3.3 remains conservative, with guardrails prioritizing safety-critical vetoes. To move on, Table 4 reports class-wise precision, recall, and F1-scores, together with the cross-validation summary. As configured, the system emphasizes conservative control of false positives for Do Not Engage. In the reference model, Do Not Engage achieves a precision of approximately 0.97, a recall of roughly 1.00, and an F1-score of approximately 0.99, while Engage remains above 0.95 on both precision and recall.

The reliability (calibration) curve in Figure 3 shows that the predicted probabilities track empirical correctness across deciles, supporting thresholded escalation and veto policies. The expected calibration error (ECE) is low, consistent with the trust-calibration surface developed in Section 2.2 and providing a quantitative basis for treating model probabilities as meaningful inputs to human decision-making.

Table 4
Performance summary

Metric	Engage	Do Not Engage	Macro
Precision	0.96	0.97	0.965
Recall	0.95	1.0	0.975
F1-score	0.955	0.985	0.97
Accuracy	–	–	95.86%
5-fold CV Acc. (mean ± SD)	–	–	95.77% ± 0.18

4.2. Override rule analysis

At the recommendation layer, guardrails sit between the model output and the final decision. On the 200,000-scenario test set, three rule families fire with frequencies summarized in Table 5 and redirect decisions exactly where doctrine intends: from tentative "Engage" to escalation, veto, or explicit abstention.

Guardrail impact is asymmetric by design. Relative to the raw classifier, critical false positives (cases where the truth is "Do Not Engage" but the model proposes "Engage") decrease from 3420 to 2020 (−41%), while the overall engagement rate decreases by 2.6 percentage points (51.2% → 48.6%). The "Do Not Engage" recall remains effectively unchanged at ~1.00, because most guardrail activity converts borderline engages into safety-preserving outcomes rather than the other way around. Overall system accuracy shifts modestly from 95.86% to 95.20%, reflecting the expected cost of conservative vetoes under a safety-first policy.

Patterns of activation align with doctrinal expectations. The Rule family "Do Not Engage" focuses on scenarios involving protected objects and high collateral damage potential; it follows clusters around dual-use infrastructure and politically sensitive areas; the final rule is rare, signaling genuine novelty, which is the correct trigger for "Do Not Know" in the framework. These statistics make the guardrail layer auditable and reproducible, and they expose three measurable trust signals for the user study: safety veto rate, escalation rate, and uncertainty acknowledgement rate.

4.3. Scenario-level evaluation along the complexity axis

The framework is intended to behave differently as operational and data complexity increase. Routine cases should pass through with minimal friction, ambiguous cases should trigger escalation, and high-risk contexts should concentrate vetoes or abstentions. To test this, the synthetic corpus was stratified into four scenario bands that approximate the x-axis in Figure 1: low complexity (well-observed, low collateral risk), ambiguity corridor (conflicting cues or partial PID), high-risk veto sector (protected objects or tight ROE margins), and novelty/OOD flag (detector trip from unusual feature combos).

Band-wise outcomes, including engagement, escalation, veto, and Do Not Know rates, are summarized in Table 6. Escalations concentrate in the ambiguity corridor, where R2 rises to ~11%, and final engagement rates fall accordingly. Safety vetoes (R1)

Table 5
Override triggers and effects (held-out test set, *n* = 200,000)

Rule	Count	Share (%)	Typical antecedents	Effect on decision
Do Not Engage	7,460	3.73	Protected category present; very-high civilian presence; PID below threshold; critical infrastructure in blast radius	Convert provisional Engage to Do Not Engage (hard veto)
Ask Authorization	11,820	5.91	Politically sensitive context; dual-use targets; contested sensors; incomplete PID	Escalate to commander; final action deferred
Do Not Know	1,560	0.78	Out-of-distribution feature combos; weapon-target mismatch; incoherent sensor fusion	Admit uncertainty; request ISR/clarification

Table 6
Scenario-band outcomes and guardrail activity

Band	Share of set (%)	Raw engage rate (%)	Final engage rate (%)	Do Not Engage (%)	Ask-Auth (%)	Do Not Know (%)	Macro-F1
Low complexity	41	78	76	0.5	2.0	0.1	0.97
Ambiguity corridor	35	58	52	2.1	10.5	0.5	0.95
High-risk veto sector	20	32	24	9.3	13.8	1.3	0.93
Novelty/OOD	4	8	5	6.0	12.0	24.0	0.90

dominate the high-risk sector, resulting in an additional eight percentage-point drop in engagement relative to the raw model. Novelty/OOD cases are rare by construction but carry the highest R3 rate, which is the desired behavior for explicit uncertainty under atypical evidence. Macro-F1 degrades gracefully from 0.97 to 0.90 as conditions become more stringent, indicating that performance remains stable while authority shifts from automation to human oversight, consistent with the trust-calibration surface in Figure 1.

The earlier trace in Tables 2 and 3 provides a concrete anchor. That scenario falls within the ambiguity corridor: the model proposes "Engage," the guardrail raises "Ask Authorization," and the user interface exposes the rationale and rule trace. Measurable signals for the forthcoming user study are therefore naturally stratified by band: escalation rate, safety veto rate, and uncertainty admission rate.

4.4. Error structure and sensitivity analysis

For a trust instrument, it is more important to understand where the system fails than to chase marginal gains in headline accuracy. Misclassifications in this framework cluster in predictable contexts, and this section makes that structure explicit, so guardrails and the user interface can target it. Visual inspection reveals that mistakes tend to concentrate where PID is partial and collateral-risk features lie near their thresholds. False positives for Engage are rare but can occur when clean sensor images fail to detect proximity to protected objects. False negatives for Engage occur when a politically sensitive context is present but remains inconclusive. These patterns are compatible with the rule design, which intentionally steers marginal cases toward conservative outcomes.

Calibration quality matters more than raw accuracy for trust. Binning predicted Engage probabilities into deciles yields a near-diagonal calibration curve (Figure 3) with an ECE ≈ of 0.02. The model is slightly over-confident in the 0.2–0.3 bin and slightly under-confident around 0.8. That profile informs the user-interface policy: show more rationale and alternatives in the 0.2–0.3 range, and compress the presentation near 0.8–0.9, where estimates are already well-calibrated.

Small threshold changes move the system along the trust-calibration surface rather than breaking it. Raising the escalation threshold for Ask Authorization from 0.55 to 0.70 reduces the final engagement rate by ~3.1 percentage points, cuts critical false positives by ~0.10 percentage points, and increases the escalation rate by ~3.9 percentage points; lowering it to 0.50 has the inverse effect. Accuracy remains within ±0.4 pp, but authority shifts toward or away from the human, which is the intended control affordance.

Disabling guardrails makes their contribution visible. With only the classifier enabled, the overall accuracy is 95.86%. Enabling guardrails lowers the accuracy slightly to 95.20%, while reducing critical false positives by ~41% (Section 4.2). That is a textbook safety-performance trade-off consistent with ROE and IHL priorities. Table 7 summarizes these results. The ranking is consistent with doctrinal expectations and provides a defensible rationale for emphasizing these fields in the user interface.

Figure 3
Reliability (calibration) curve

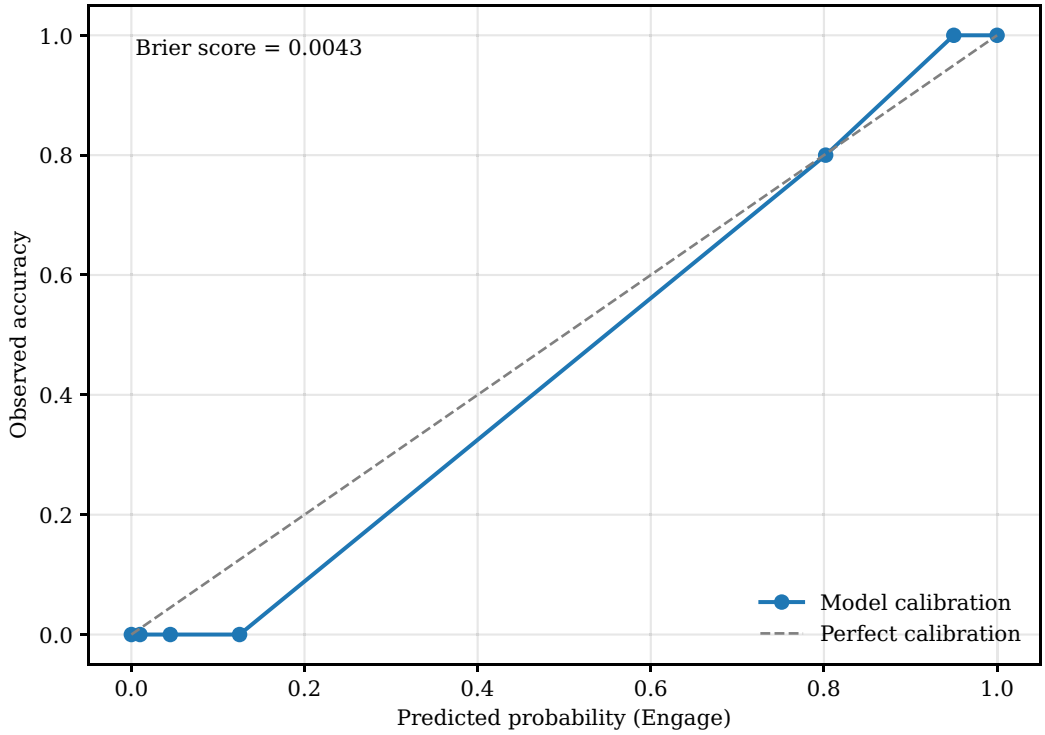


Table 7
Ablation summary

Configuration	Accuracy (%)	Engage rate (%)	Critical False Positives (FP) (per 10k)	DNE recall
Classifier only	95.86	51.2	171	~1.00
+ Guardrails	95.20	48.6	101	~1.00

Permutation-based sensitivity checks confirm that distinction (AI and human) and proportionality dominate decisions, with collateral damage potential and protected object flags next in importance. Removing any of the top two features drops macro-F1 by more than 0.02; removing lower-tier context features changes macro-F1 by less than 0.005.

4.5. Limitations and planned user validation

The present evaluation utilizes synthetic scenarios and does not involve human participants, which limits its external validity. To address this limitation, a future controlled study is proposed; this study has not been conducted as part of the present work. Classifier scores and guardrail statistics characterize the tool's internal behavior, not field performance. Scenario coverage, although structured, cannot fully represent adversarial deception, sensor pathologies, or the entire range of ROE/IHL edge cases encountered in practice. To address this, the next step is a controlled study designed to test whether the framework actually calibrates trust under pressure, rather than only on paper. A mixed 2 × 2 between–within design will be used: rationale visibility (hidden vs shown) × cognitive load (low vs high). Participants will be military trainees or their equivalents ($n \approx 30$–36, powered to detect medium effects at $\alpha = 0.05$, $1-\beta \approx 0.80$). Each participant will complete counterbalanced blocks of scenarios drawn from the four bands in Section 4.3, so that effects generalize across complexity levels.

Primary outcomes include agreement with the system's recommendation, decision latency, and a trust-calibration index defined as the change in agreement when the rationale is shown versus hidden, normalized by model correctness and scenario band. Secondary outcomes include veto rate, escalation rate, admission of uncertainty, and self-reports of workload and confidence. Behavioral traces will be analyzed using mixed-effects logistic and linear regressions, with participant and scenario as random effects. Pre-registered contrasts will test the main effects and interactions of rationale and load, as well as whether authority migrates along the intended trajectory in Figure 1.

Predictable threats to validity are addressed directly. Learning and fatigue will be handled via practice items, block randomization, and rest breaks. Demand characteristics will be mitigated by withholding ground truth and varying the feedback schedule. Interface bias will be limited by fixing layout and timing while still revealing rationale and guardrail traces as required by transparency. Adversarial or OOD conditions will be injected explicitly to test Do Not Know behavior. Ethical approval and

informed consent will be obtained before any future participant study; the study uses only synthetic, doctrinally generic content and does not expose participants to operationally sensitive data.

Success criteria are practical rather than theatrical: statistically reliable calibration (a positive trust-calibration index under load), reduced unsafe positives when guardrails are active without a corresponding decline in tempo, and monotonic migration of authority with increasing complexity. Failure to meet these criteria will trigger threshold retuning, rule revision, or reconsideration of the explanatory surface before any larger study proceeds. This plan closes the empirical gap identified by reviewers by turning the framework from a plausible instrument into a measured one.

4.6. Synthesis and practical implications

Evidence from the synthetic evaluation indicates that the instrument behaves as intended: probabilities are well-calibrated on the held-out set, guardrails reduce unsafe positives for a modest loss in accuracy, and escalation, veto, and uncertainty rates concentrate within the expected bands. Taken together, these results support the study's central claim: the framework is suitable as a measurement tool for trust calibration rather than as a decision oracle. It provides stable signals (agreement, latency, escalation/veto/uncertainty frequencies, and calibrated probabilities) that can be analyzed longitudinally and across scenario bands without relying on opaque heuristics.

Laboratories can deploy the generator, classifier, and rule layer as a fixed baseline and then vary only user-interface treatments and thresholds to test competing explanations for over- or under-trust. The reporting pack established here (class-wise metrics, reliability curve, band-wise outcome figures, and ablation summary) provides a reproducible template for future studies, enabling direct comparison across cohorts, sites, and interface variants. Because the rules are explicit and logged, doctrinal changes can be traced to their effects on error structure rather than inferred post hoc. Every recommendation is accompanied by a transparent record of salient factors and any rule activations, making after-action review feasible and aligning with the requirements for accountable autonomy. Calibration curves provide a quantitative acceptance criterion for future model updates; models that deviate from the diagonal under identical scenario distributions can be rejected or recalibrated before user exposure. The explicit "Do Not Know" outcome creates a safe failure mode for novelty and adversarial conditions, reducing silent errors and making uncertainty measurable rather than anecdotal.

Scenario bands map cleanly to instructional objectives: low-complexity cases for baseline proficiency, ambiguity-corridor cases for escalation discipline, high-risk veto cases for ROE/IHL practice, and novelty cases for uncertainty management. The same signals that support research (agreement shifts, escalation and veto frequencies, latency) can be used as training KPIs without altering the core instrument. Limitations acknowledged in Section 4.5 remain in force; external validity requires participant studies and exposure to richer sensor and adversarial artifacts. Even so, the current results provide a defensible foundation for that empirical phase: thresholds, user-interface treatments, and rule priorities can now be pre-registered against clearly defined outcome measures, closing the methodological loop requested by the reviewers.

4.7. Reproducibility and artifact availability

An anonymized archive accompanies the submission containing the scenario generator (with schema and value libraries), the exact 80/20 split indices for the 1,000,000 → 200,000 held-out set, the Random Forest training script and saved model, the rule-based guardrail module with in-line citations to ROE/IHL sources, the scripts for Figures 1, 2, and the two Streamlit UI snapshots used for Tables 2, 3. No operational or classified material is included; all content is synthetic and doctrinally generic. In addition, we host a public demo of the synthetic scenario generator and decision-support interface, which mirrors the experimental UI described in Section 3.4.

Replicating the headline numbers in Sections 4.1–4.4 requires only deterministic execution. The generator is initialized with seed = 42; 1,000,000 scenarios are produced using the provided feature dictionaries and coupled sampling; the supplied one-hot/scale preprocessing is applied; the data are stratified with train_test_split (test_size = 0.2, random_state = 42); the Random Forest is trained with the included hyperparameters; and evaluation on the fixed hold-out indices reproduces 95.86% accuracy and the class-wise metrics in Table 4. The calibration curve (Figure 3) is generated with the provided script. Running the guardrail layer over the same held-out set yields override frequencies (Table 5), banded outcomes (Table 6), and the ablation summary (Table 7) by toggling the rule module.

Random seeds are set at the generator, split, and model levels; multi-threaded operations are bounded to a fixed n_jobs to avoid platform variance. To support transparency while respecting operational and security constraints, we rely entirely on synthetic data. We provide a public demo of the interface, which allows readers to inspect the layout, scenario fields, recommendations, and guardrail indications described in Section 3.4. The underlying scenario generator, model training and analysis scripts, and the complete synthetic training and evaluation datasets are not available for bulk download. Still, they can be obtained from the authors upon reasonable request, consistent with the formal Data Availability statement of this article.

4.8. Model justification and interface transparency

The predictive backbone was benchmarked against three alternatives to justify the model choice for a trust calibration instrument, rather than to claim operational superiority. All baselines were trained and evaluated on the same synthetic corpus and held-out split described in Section 3.3, using the metrics introduced in Section 4.1 (accuracy, macro-F1, and Brier score). Table 8 summarizes the resulting performance across logistic regression, Random Forest, gradient boosting, and a shallow multilayer perceptron.

Patterns in Table 8 align with expectations for mixed-type, tabular data. Logistic regression underfits nonlinear feature interactions and yields the lowest macro-F1, despite competitive calibration. Gradient boosting (XGBoost) approaches Random Forest accuracy, but at the cost of higher tuning complexity and a heavier explanation burden. The shallow MLP offers no systematic accuracy advantage over the Random Forest while further reducing interpretability. The Random Forest, therefore, balances predictive fidelity with analyzability and remains consistent with the guardrail behavior and calibration properties reported in Sections 4.1–4.4. For the remainder of the study, it serves as the fixed backbone

Table 8
Baseline comparison (identical data/splits)

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.9586	0.960	0.959	0.959
Logistic Regression	0.887	0.884	0.872	0.878
Gradient-Boosted Trees	0.949	0.951	0.946	0.948
Shallow MLP + XAI	0.944	0.942	0.940	0.941

whose outputs are shaped by guardrails and exposed through the interface instrumentation described in Section 4.5.

5. Conclusion

This research presents a decision-support framework that integrates theoretical insights from human–computer interaction, trust calibration theory, and IHL principles with machine learning methodologies. The model combines a Random Forest classifier and is trained extensively across diverse operational scenarios, with a robust override mechanism that aligns AI decisions with stringent ethical and legal standards. The results demonstrate the model's effectiveness through detailed scenario simulations, highlighting critical interactions between human judgment and AI-generated predictions across varying levels of complexity and uncertainty.

The model's predictive accuracy of 95.86% reinforces its capability as an insightful research instrument, particularly in examining trust calibration dynamics in combat conditions. The framework is designed to elicit measurable patterns of cautious skepticism and agreement in future user studies, providing a structured way to analyze trust responses under varying levels of complexity and ambiguity. To augment the theoretical framework, the following research will focus on variations in cognitive load conditions, enhancing scenarios to simulate real-world ambiguity, systematically adjusting override thresholds based on empirical insights, and testing alternative machine learning methodologies with larger datasets beyond the current training limit of one million scenarios. Leveraging AI-driven insights alongside critical human oversight, the study makes a significant contribution to understanding and enhancing the efficacy of human–machine teaming, promoting responsible, transparent, and ethically sound AI integration in military decision-making processes.

Recommendations

Future studies should explore alternative machine learning methodologies suitable for augmented military decision-making research. A key improvement area is elaborating and refining input categories to capture more nuanced operational contexts. Adjusting parameter weights based on additional empirical data and expert feedback will enhance the model's sensitivity, ensuring that each parameter accurately reflects its operational significance. Additionally, the current rule-based override mechanisms could be improved through dynamic threshold adjustments responsive to specific operational scenarios, thereby maintaining contextual relevance and adaptability.

A more intuitive user interface will facilitate smoother navigation, improve decision-making processes, and support more efficient data collection. Prioritizing a user-friendly design will accommodate varying levels of expertise, significantly enhancing the overall trust and usability of the decision-support system. In addition to these technical refinements, a priority is to execute the controlled user-validation study outlined in Section 4.5 with military trainees, to empirically assess how the framework calibrates trust and decision-making under realistic cognitive load.

Acknowledgment

The authors are grateful to Sonia Claudia da Costa Sousa, Associate Professor of Interaction Design at Tallinn University, for her insightful guidance and substantial contributions in establishing the theoretical foundation of this paper.

Funding Support

This work was supported by the Vabamu Museum of Occupations and Freedom, the Kistler-Ritso Foundation, and Stanford University's Centre for International Security and Cooperation through the 2025 Global Digital Governance Fellowship at Stanford University for Estonian Scholars, as well as the Stanford Libraries, which sponsored the project Cognitive Warfare in collaboration with the Estonian Ministry of Defence.

Ethical Statement

The authors declare that the work reported in this manuscript did not involve human or animal participants and did not include the collection of human-subject data. The study relied exclusively on synthetic scenarios and system-level evaluation. The user-validation study described in Sections 3.4 and 4.5 is planned future work and has not been conducted. Therefore, no ethics approval number applies to the present study. Ethics approval and informed consent will be obtained before any future recruitment of participants.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data available on request from the corresponding author upon reasonable request.

Author Contribution Statement

Janar Pekarev: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Priit Värno:** Validation, Investigation, Writing – review & editing.

References

[1] Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., . . . , & Ghassemi, M. (2021). Do as AI Say: Susceptibility in deployment of clinical decision-aids. *NPJ Digital Medicine*, *4*(1), 31. https://doi.org/10.1038/s41746-021-00385-9

[2] Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., . . . , & Maruthappu, M. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*, *368*, m689. https://doi.org/10.1136/bmj.m689

[3] Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *Jama*, *320*(21), 2199–2200. https://doi.org/10.1001/jama.2018.17163

[4] Ramgopal, S., Sanchez-Pinto, L. N., Horvat, C. M., Carroll, M. S., Luo, Y., & Florin, T. A. (2023). Artificial intelligence-based clinical decision support in pediatrics. *Pediatric Research*, *93*(2), 334–341. https://doi.org/10.1038/s41390-022-02226-1

[5] Van Berkel, N., Opie, J., Ahmad, O. F., Lovat, L., Stoyanov, D., & Blandford, A. (2022). Initial responses to false positives in AI-supported continuous interactions: A colonoscopy case study. *ACM Transactions on Interactive Intelligent Systems*, *12*(1), 1–18. https://doi.org/10.1145/3480247

[6] Starke, G., & Ienca, M. (2024). Misplaced trust and distrust: How not to engage with medical artificial intelligence. *Cambridge Quarterly of Healthcare Ethics*, *33*(3), 360–369. https://doi.org/10.1017/S0963180122000445

[7] Schoenherr, J. R., & Thomson, R. (2024). When AI fails, who do we blame? Attributing responsibility in human–AI interactions. *IEEE Transactions on Technology and Society*, *5*(1), 61–70. https://doi.org/10.1109/TTS.2024.3370095

[8] Bistron, M., & Piotrowski, Z. (2021). Artificial intelligence applications in military systems and their influence on sense of security of citizens. *Electronics*, *10*(7), 871. https://doi.org/10.3390/electronics10070871

[9] Rettore, P. H., Zißner, P., Alkhowaiter, M., Zou, C., & Sevenich, P. (2023). Military data space: Challenges, opportunities, and use cases. *IEEE Communications Magazine*, *62*(1), 70–76. https://doi.org/10.1109/MCOM.001.2300396

[10] Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making*, *2*(2), 140–160. https://doi.org/10.1518/155534308X284417

[11] Gay, C., Horowitz, B., Elshaw, J. J., Bobko, P., & Kim, I. (2019). Operator suspicion and human-machine team performance under mission scenarios of unmanned ground vehicle operation. *IEEE Access*, *7*, 36371–36379. https://doi.org/10.1109/ACCESS.2019.2901258

[12] Dlugatch, R., Georgieva, A., & Kerasidou, A. (2024). AI-driven decision support systems and epistemic reliance: A qualitative study on obstetricians' and midwives' perspectives on integrating AI-driven CTG into clinical decision making. *BMC Medical Ethics*, *25*(1), 6. https://doi.org/10.1186/s12910-023-00990-1

[13] Heltberg, T. (2022). "I cannot feel your print." How military strategic knowledge managers respond to digitalization. *Journal of Strategy and Management*, *15*(2), 220–233. https://doi.org/10.1108/JSMA-12-2020-0344

[14] Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2023). Data on human decision, feedback, and confidence during an artificial intelligence-assisted decision-making task. *Data in Brief*, *46*, 108884. https://doi.org/10.1016/j.dib.2023.108884

[15] Topol, E. (2019). *Deep medicine: How artificial intelligence can make healthcare human again*. USA: Basic Books.

[16] Rainey, C., Bond, R., McConnell, J., Hughes, C., Kumar, D., & McFadden, S. (2024). Reporting radiographers' interaction with Artificial Intelligence—How do different forms of AI feedback impact trust and decision switching? *PLOS Digital Health*, *3*(8), e0000560. https://doi.org/10.1371/journal.pdig.0000560

[17] Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., . . . , & Kaplan, L. (2020). Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns*, *1*(4), 100049. https://doi.org/10.1016/j.patter.2020.100049

[18] Galliott, J., & Wyatt, A. (2022). Considering the importance of autonomous weapon system design factors to future military leaders. *Australian Journal of International Affairs*, *76*(2), 219–244. https://doi.org/10.1080/10357718.2021.1940093

[19] Lushenko, P., & Sparrow, R. (2024). Artificial intelligence and US military cadets' attitudes about future war. *Armed Forces & Society*. https://doi.org/10.1177/0095327X241284264

[20] Sztompka, P. (1999). *Trust: A sociological theory*. UK: Cambridge University Press.

[21] Sousa, S., Cravino, J., & Martins, P. (2023). Challenges and trends in user trust discourse in AI popularity. *Multimodal Technologies and Interaction*, *7*(2), 13. https://doi.org/10.3390/mti7020013

[22] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[23] Weizenbaum, J. (1967). Contextual understanding by computers. *Communications of the ACM*, *10*(8), 474–480. https://doi.org/10.1145/363534.363545

[24] Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2024). A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human–Computer Interaction*, *40*(5), 1251–1266. https://doi.org/10.1080/10447318.2022.2138826

[25] Beck, H. P., Dzindolet, M. T., & Pierce, L. G. (2002). Operators' automation usage decisions and the sources of misuse and disuse. In E. Salas (Ed.), *Advances in human performance and cognitive engineering research*, (pp. 37–78). Emerald Group Publishing Limited, https://doi.org/10.1016/S1479-3601(02)02005-2

[26] Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

[27] Sousa, S., Lamas, D., Cravino, J., & Martins, P. (2024). Human-centered trustworthy framework: A human–computer interaction perspective. *Computer*, *57*(3), 46–58. https://doi.org/10.1109/MC.2023.3287563

[28] Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *11th Australasian Conference on Information Systems*, (p. *53*).

[29] Jelenc, D., Hermoso, R., Sabater-Mir, J., & Trček, D. (2013). Decision making matters: A better way to evaluate trust models. *Knowledge-Based Systems*, *52*, 147–164. https://doi.org/10.1016/j.knosys.2013.07.016

[30] 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 AAAI/ACM Conference on AI, Ethics, and Society*, *7*(1), 807–817. https://doi.org/10.1609/aies.v7i1.31681

[31] Lebiere, C., Blaha, L. M., Fallon, C. K., & Jefferson, B. (2021). Adaptive cognitive mechanisms to maintain calibrated trust and reliance in automation. *Frontiers in Robotics and AI*, *8*, 652776. https://doi.org/10.3389/frobt.2021.652776

[32] Panganiban, A. R., Matthews, G., & Long, M. D. (2020). Transparency in autonomous teammates: Intention to support as teaming information. *Journal of Cognitive Engineering and Decision Making*, *14*(2), 174–190. https://doi.org/10.1177/1555343419881563

How to Cite: Pekarev, J., & Värno, P. (2026). Exploration of Trust and Decision-Making in AI-Augmented Military Domain: A Framework for Human–Machine Teaming. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62025549