THE WALKER PAPERS

Trust but Verify

The Character, Competence, & Control of Large Language Models

Michael S. Perry, Lt Col, USAF

21

AIR UNIVERSITY

AIR FORCE FELLOWS

Trust but Verify

The Character, Competence, & Control of Large Language Models

Michael S. Perry, Lt Col, USAF

Walker Paper No. 21

Advisors:
Dr. Margaret E. Kosal, Dr. Tyler Cook & Hon. John Tien
Georgia Institute of Technology

Air University Press
Academic Services
Maxwell Air Force Base, Alabama

Acting Director, Air University Press
Dr. Stephanie Rollins

Project Editor
Dr. Achala Gunasekara-Rockwell

Illustrator
Catherine Smith

Print Specialist
Jonathan Marks

Air University Press
600 Chennault Circle, Building 1405
Maxwell AFB, AL 36112-6010
https://www.airuniversity.af.edu/AUPress

Facebook:
https://facebook.com/AirUnivPress

X:
https://X.com/aupress

LinkedIn:
https://www.linkedin.com/company/air-university-press/

Instagram:
https://www.instagram.com/air_university_presss

Accepted by Air University Press May 2025 and published March 2026.

AIR UNIVERSITY PRESS

Contents

Illustrations

Figures

Tables

Air Force Fellows

Since 1958, the Air Force has assigned a small number of carefully chosen, experienced officers to serve one-year tours at distinguished civilian institutions studying national security policy and strategy. Beginning with the 1994 academic year, these programs were accorded senior service school professional military education in-residence credit. In 2003 these fellowships assumed senior developmental education (SDE), force development credit for eligible officers.

The SDE-level Air Force Fellows serve as visiting military ambassadors to their centers, devoting effort to expanding their colleagues' understanding of defense matters. As such, candidates for SDE-level fellowships have a broad knowledge of key Department of Defense (DOD) and Air Force issues. SDE level fellows perform outreach by their presence and voice in sponsoring institutions. SDE-level fellows are expected to provide advice, promote, and explain Air Force and DOD policies, programs, and military doctrine strategy to nationally recognized scholars, foreign dignitaries, and leading policy analysts. The Air Force Fellows also gain valuable perspectives from the exchange of ideas with these civilian leaders. SDE-level fellows are expected to apprise appropriate Air Force agencies of significant developments and emerging views on defense and economic and foreign policy issues within their centers. Each fellow is expected to use the unique access she or he has as grounds for research and writing on important national security issues. The SDE Air Force Fellows include the National Defense Fellows, the RAND Fellows, the National Security Fellows, and the Secretary of Defense Corporate Fellows. The Air Force Fellows program also supports a post-SDE military fellow at the Council on Foreign Relations.

On the intermediate developmental education level, the chief of staff approved several Air Force fellowships focused on career broadening for Air Force majors. The Air Force Legislative Fellows was established in April 1995 with the Foreign Policy Fellowship and Defense Advanced Research Projects Agency Fellowship coming under the Air Force Fellows program in 2003. In 2004, the Air Force Fellows also assumed responsibility of the National Laboratories Technologies Fellows.

About the Author

Colonel Michael S. Perry is the Strategy Development Division Chief in the Joint Staff J-5, charged with providing military advice on strategy development, strategic risk assessment and combatant command organization. This includes the overall management associated with the formulation, development, and production of Joint Staff positions for the National Security Strategy (NSS), National Defense Strategy (NDS), National Military Strategy (NMS), Chairman's Risk Assessment (CRA), Unified Command Plan (UCP), and Annual Joint Assessments (AJA).

Col Perry enlisted in the Air Force in 1997 and received his commission from Officer Training School in 2004. He has spent most of his career in special operations flying the MC-130.

Prior to this assignment, he was a National Defense Fellow in the Sam Nunn School of International Affairs at Georgia's Institute of Technology.

Acknowledgments

Like many officers of my generation, I've watched artificial intelligence evolve from an abstract curiosity into a battlefield consideration. At first, it was all novelty—chess-playing machines, self-driving cars that couldn't navigate a snowstorm, and voice assistants that confused "call wife" with "play Whitesnake." But then came AlphaGo. In 2016, the world watched as DeepMind's AlphaGo defeated one of the greatest Go players in history. A milestone, yes—but still a game with perfect information. The true shift came when AI systems began outperforming humans in games where not all information is known—games like Texas Hold 'Em poker, where bluffing matters, and Diplomacy, where reading your opponent is as crucial as moving pieces on a board. These victories hinted at something far more consequential than superior calculation: they suggested machines could navigate uncertainty, deception, and ambiguity—traits once thought uniquely human.

It was around this time I began to ask myself a different question: if AI can outplay humans in strategic games of incomplete information, could it one day augment the decisions made by generals and statesmen? Could a large language model—trained not only on doctrine but on centuries of statecraft and conflict—contribute meaningfully to national security decision-making? To answer that question, I needed a new framework. Military leaders are trained to trust people (wingmen, subordinates, and commanders) but not machines. And yet, the fog of war is thickening in the age of intelligent systems. If we are to rely on LLMs to support our most difficult strategic choices, we must first learn how to assess their trustworthiness. This paper is the result of that inquiry. It aims to offer a practical approach, grounded in the logic of military trust and adapted to the peculiarities of algorithmic reasoning, to help decision-makers know when a machine is not just capable—but credible.

Abstract

As large language models (LLM) evolve from linguistic curiosities into strategic instruments, the US military must confront a critical question: when, and how, can these machines be trusted? This paper proposes a pragmatic framework for evaluating the trustworthiness of LLMs in military decision-making contexts. Borrowing from established human trust models and tailored for the algorithmic age, the "Trust Triad"—Character, Competence, and Control—offers senior military leaders a structured method for assessing LLMs intended to augment judgment, not replace it.

The analysis spans the full arc of military decision support, from data aggregation to wargaming and planning. It shows that while LLMs are already useful for accelerating routine staff tasks, their integration into more analytical and operational roles demands new standards of trustworthiness. Using weighted metrics derived from the TrustLLM evaluation suite, this paper provides a comparative assessment of current models, revealing significant differences in their ethical alignment, factual reliability, and robustness under pressure.

The conclusion is clear but not final: no model is perfect, but some are more fit for military purpose than others—and they are improving fast. The paper also identifies key gaps in current evaluation frameworks, particularly in measuring transparency and accountability. To address these, it recommends further research into standardized metrics such as the Transparency Evaluation Score and Attribution Traceability Score. Trust, in war as in technology, is earned. This paper aims to help military leaders distinguish between systems that merely perform, and those that are worthy of command confidence.

Introduction . . . Trust, but Verify

Technology is a useful servant but a dangerous master.

—Christian Lous Lange

If there is one technology capable of ushering in a new era of American military dominance, it is artificial intelligence. Specifically, large language models (LLM)—those eerily eloquent engines of reasoning and recall—have rapidly evolved from mere text generators into plausible partners for human judgment. They are already adept at brainstorming,[1] summarizing dense documents,[2] and extracting actionable insights from terabytes of unstructured data.[3] But something more consequential is afoot: LLMs are now encroaching upon decision-support roles, helping users not just find facts, but synthesize them into analysis and, increasingly, advice.[4] This pivot demands a new kind of scrutiny.

For commanders and senior military leaders, the question is no longer whether LLMs are useful (they plainly are) but whether they can be trusted. And trust, as any field officer will confirm, is a fragile currency in wartime. Trust must be earned, not through marketing brochures or vague promises of responsible AI, but through measurable, field-tested performance under pressure. The fog of war, once described as the enemy of clarity, now comes laced with metadata. In this modern battlespace, LLMs offer the potential to cut through the noise—surfacing relevant intelligence, simulating enemy behavior, and generating well-reasoned plans. Indeed, LLMs can assist at nearly every phase of the Joint Planning Process:

Initiation & Mission Analysis: Summarize and structure strategic guidance, operations orders (OPORD), and intelligence documents to extract mission-essential tasks and constraints and speed up initial understanding and alignment with higher headquarters' intent.

COA Development: Generate multiple course-of-action (COA) options based on predefined objectives, constraints, and terrain data to stimulate creativity and help planners consider unorthodox or underexplored approaches.

COA Analysis (Wargaming): Simulate opposing force actions and outcomes using historical analogs or probabilistic models to offer additional perspectives or red-team critiques, testing COA robustness.

COA Comparison: Assist in creating decision matrices, comparing COAs against criteria (feasibility, acceptability, suitability, etc.) to enable rapid revision of weighting schemes and visualizations for senior leader decisions.

Plan or Order Development: Draft annexes, appendices, and fragments of OPORDs based on prior examples and current data to automate low-value but time-intensive writing tasks, allowing staff to focus on strategy.

Execution Planning & Transition: Track task completion, generate situation reports (SITREP), and flag deviations from planned timelines to enhance battle rhythm and continuity during staff handovers.

These capabilities are no longer speculative. They are in development or deployment across various government and commercial labs. The Trump administration's push to transition AI from experimentation to maturation means the Pentagon will soon adopt these tools at scale.[5] The question for commanders is no longer technical, but moral and operational: When should I trust this machine? And when should I not?

Here lies the central paradox of LLMs: they are brilliant, but opaque. LLM judgments are probabilistic, and their reasoning is untraceable. This inability to understand their outputs forces us to extend trust. This report proposes a solution: the Trust Triad—character, competence, and control. These pillars reflect the criteria military leaders already use to assess human subordinates. Character addresses the model's ethical compass—does it align with military norms and values? Competence measures its performance under fire—can it remain accurate and coherent in novel, adversarial, or chaotic settings? And control captures whether humans remain meaningfully in the loop—do commanders retain the power to override, audit, and understand the machine's output?

In the pages ahead, this paper will treat LLMs as junior lieutenants—flawed, capable, and in need of constant evaluation. Using a field-adapted evaluation framework, it will map the terrain of trustworthiness: identifying which models are mission-ready, which are ethically brittle, and which might fail at the worst possible moment. There is no doctrine yet for fighting wars with LLMs at your elbow. But if one is to be written, it must begin with trust—not trust as a sentiment, but as a metric.

Applying LLM's . . . Across the Decision-Making Landscape

Machines don't fight wars. People do. And they use their minds.

—Col John Boyd

Understanding how humans make decisions is a necessary first step to evaluating how LLMs might one day contribute to strategic military planning. If LLMs are to assist in decision-making, military leaders must first recognize the structures and limitations of their own processes. Humans, it turns out,

employ a range of decision-making approaches—some rational, others intuitive, and often shaped by the constraints of time, pressure, and available information. The Pentagon is a place where decisions range from the granular to the grand strategic. Each of these calls falls into a broad taxonomy of decision-making styles that scholars and practitioners alike have long studied.

The first major distinction in decision-making is between rational, structured choices and those guided by intuition. Rational decision-making, a hallmark of bureaucracies, involves systematic analysis, data collection, and explicit evaluation of alternatives.[6] The Department of War (DOW) thrives on this approach when conducting force structure assessments or weapons procurement.[7] Think of the McNamara-era embrace of systems analysis: a logic-driven approach that, in theory, reduced national security decision-making to an equation. An LLM can support these types of decisions by rapidly synthesizing vast datasets and generating alternative courses of action to aid force structure assessments or weapons procurement.

Yet, rationality has its limits. It is slow, often overly reliant on questionable assumptions, and frequently impractical in the real world.[8] This is where intuitive decision-making comes in. Pilots in combat do not conduct Bayesian probability assessments before taking evasive action. Instead, they rely on pattern recognition and experience.[9] Studies of decision-making under stress, such as Gary Klein's work on fireground commanders, suggest that experts use intuition to rapidly assess familiar situations, drawing on vast mental libraries of past experiences.[10] Military leaders, especially in combat, have long understood that some of the best decisions are made without time for careful deliberation. LLMs can augment intuitive decision-making by surfacing relevant patterns from historical data and analogous scenarios in real time, effectively expanding a leader's mental library to support faster, experience-informed judgments under pressure.

Another defining characteristic of decision-making is whether it is made alone or collectively. The US military is no stranger to the power and perils of group decision-making. The Joint Chiefs of Staff, the National Security Council, and countless working groups bring together minds from different backgrounds to reach consensus. This method has undeniable benefits: diverse perspectives, collective expertise, and institutional legitimacy. But group decision-making also has well-documented flaws. "Groupthink," famously outlined in studies of the Kennedy administration's handling of the Bay of Pigs invasion, demonstrates the risks of excessive consensus.[11] Military leaders are trained to speak up in planning sessions, but rank, culture, and organizational inertia often mean that dissenting voices are drowned out.[12] LLMs,

in theory, could act as devil's advocates—offering alternative perspectives and challenging flawed assumptions.

The luxury of time is rarely afforded in military decision-making. Some choices, such as wargame simulations for future conflicts, can be deliberated for years. Others—like responding to a cyberattack or intercepting a ballistic missile—must be made in minutes or seconds.[13] The OODA loop (Observe, Orient, Decide, Act), developed by Air Force Colonel John Boyd, captures the importance of speed in modern warfare.[14] Leaders who can cycle through decisions faster than their adversaries gain an edge. It is no wonder that many military commanders favor 80 percent solutions delivered quickly over theoretically perfect answers that arrive too late. LLMs could accelerate decision-making by processing vast amounts of information faster than any human could.[15] However, speed alone is insufficient. A flawed decision made instantaneously is still a flawed decision. The challenge lies in ensuring that LLM-generated recommendations are not just fast but also reliable and aligned with strategic goals.

Military decisions, particularly at the strategic level, are not made in a vacuum. Ethical considerations shape choices at every level, from targeting decisions in drone warfare to the trade-offs between deterrence and escalation in nuclear strategy.[16] The application of LLMs in military decision-making will require a deep understanding of these ethical dimensions.[17] Should a model prioritize minimizing civilian casualties over maximizing mission success? Should it be programmed to factor in long-term geopolitical consequences, or merely optimize for short-term tactical gains?

The use of LLMs in decision support will also face institutional and political hurdles. The DOW has long relied on established hierarchies and chains of command. Introducing LLMs into the mix disrupts traditional models of authority.[18] Who takes responsibility if an LLM-generated recommendation leads to strategic failure? Military leaders will need clear frameworks for accountability before embracing LLM-driven decision-support systems.

For LLMs to play a role in decision-making, they must first be trusted. And for them to be trusted, they must align with the way humans make decisions—not just in terms of logic, but also in how they navigate uncertainty, pressure, and ethical trade-offs. The military does not need LLMs that merely regurgitate data; it needs models that understand context, weigh competing priorities, and integrate into existing decision-making structures.

Understanding the taxonomy of human decision-making is a prerequisite for designing LLMs that will be useful rather than disruptive. The question is not whether LLMs will be used in military decision-making—the pace of technological advancement and political drive makes that inevitable.[19] The

question is whether military leaders will be ready to integrate them effectively when the time comes. That, in itself, is a decision worth making carefully.

Interpersonal & Interfacial Trust . . . From Lieutenants to LLMs

Trust is the bedrock of our profession.

—US Army Doctrine

Understanding human trust is foundational to grasping the complexities of trust a human can place in an LLM. Research across philosophy, psychology, and sociology offers insights into how people trust each other (interpersonal trust).[20] Although traditional interpersonal trust models do not map perfectly onto human-machine interactions, reframing them in this way provides a more practical and insightful lens than relying solely on machine-specific frameworks.[21]

I define the confidence a human places in an LLM to act as a reliable, ethical, and mission-aligned partner as *interfacial trust*. The term draws from the concept of an interface—the point where two distinct systems interact—highlighting the unique trust relationship formed not between peers, but between a human and a machine. Just as interpersonal trust defines confidence between people, interfacial trust captures the emerging trust dynamic at the boundary between human cognition and machine intelligence.

Evaluating LLMs

Integrating powerful LLM tools into military decision support demands a critical and nuanced understanding of their inherent limitations and potential pitfalls. Just as the military rigorously evaluates the performance characteristics of conventional weaponry, so too must we scrutinize the reliability and integrity of these algorithmic assets. Researchers are pushing beyond standard evaluation metrics, developing innovative algorithmic approaches to assess LLMs with greater depth and precision. One such method, LLMMaps, enables stratified evaluation, offering a granular breakdown of model performance across different task categories. This helps pinpoint a model's strengths and weaknesses across multiple dimensions.[22] Competitive benchmarking tools, like the LangTest[23] and TrustLLM Leaderboards,[24] provide ongoing assessments to track model improvements over time. These advancements offer a more comprehensive understanding of LLM capabilities, moving beyond surface-level performance metrics to a structured, multidimensional evaluation framework.[25]

But these frameworks overlook a key understanding of trust: it is contextual. For example, I trust my dog not to eat food off the table while I am in the room. But without the context of accountability, the trust I put in my dog fades. Similarly, LLMs shown to be competent in the areas of research and development may not maintain their high marks in combat.

Thus, the disparate benchmarks and evaluations of the current field do not adequately account for the unique ethical and operational stakes inherent in defense applications.[26] Existing evaluations may not sufficiently prioritize the alignment of LLM outputs with US strategic aims or ensure consistent reliability under the dynamic and high-stakes conditions of military operations.[27] Furthermore, the imperative of maintaining unambiguous human control within the chain of command remains a secondary consideration in general-purpose LLM evaluations.[28] This is why military leaders need a new framework to understand interfacial trust, and that framework should be grounded in our understanding of interpersonal trust.

Interpersonal and Interfacial Trust

The Pentagon has long relied on trust to keep its war machine running smoothly. Pilots trust their wingmen, commanders trust their intelligence analysts, and supply sergeants (sometimes) trust that the paperwork is actually in order. Now, as the military experiments with LLMs to assist in decision-making, a fundamental question looms: How do we know when to trust these algorithms? Rather than reinvent the wheel, the military should take a page from what it already knows best—trust frameworks developed for human relationships. Applying these models to LLMs might seem counterintuitive, like saluting a chatbot, but there is a solid logic behind it.

Several theoretical frameworks explore the nature of interpersonal trust. One perspective views trust as being composed of distinct dimensions, such as cognition-based trust and affect-based trust.[29] Cognition-based trust relies on judgments about another's competence, reliability, and integrity, often based on observable behaviors and past interactions.[30] This type of interpersonal trust is most relevant to interfacial trust. Affect-based trust builds on cognitive trust and stems from emotional bonds, feelings of goodwill, and positive regard for the other person.[31]

This type of interpersonal trust can develop between human and machine when the user inappropriately anthropomorphizes the machine. Understanding these dimensions of interpersonal trust, especially how cognitive trust aligns with interface-level trust and how affective trust can be misapplied,

helps clarify why users may develop misplaced confidence in machines like LLMs, with serious implications for decision-making.

While traditional models of interpersonal trust cannot be directly applied to LLMs due to fundamental differences between humans and machines,[32] the core constructs and influencing factors of human trust offer a valuable conceptual foundation for military decision-makers.[33] Human trust frameworks have been rigorously developed over decades, giving military leaders a structured way to evaluate trust in people. These frameworks measure intent, integrity, capability, and results—the very same dimensions that determine whether a lieutenant will get promoted or find himself stuck counting rations in a supply depot.[34] LLMs, though distinctly nonhuman, must still earn trust before they are widely adopted in mission-critical roles.[35] If these models are to support decision-making, they need to be judged by similar criteria.

One of the great paradoxes of modern LLMs is that, while they can analyze more data in a second than a human could in a lifetime, their process remains opaque. Military leaders do not—and should not—trust things they cannot understand.[36] Here, human trust frameworks help. These models already account for complex and unpredictable behavior in human relationships, offering a guide for managing the uncertainty that comes with LLM decision-making.

Future battlefields may be defined by seamless human-AI collaboration. If an officer must rely on an LLM to recommend resource allocations, generate risk assessments, or synthesize intelligence reports, they need to trust that the model is both reliable and free of hidden biases.[37] Human trust frameworks provide an essential playbook for developing protocols, training, and best practices that foster productive human-machine teams.

The military has a responsibility to deploy AI systems ethically and responsibly. Human trust frameworks help ensure that these systems are not just technically sound but also aligned with military values. After all, an LLM might be perfectly capable of summarizing a war plan, but can it be trusted to do so without reinforcing historical biases or generating misleading outputs? Human trust models help define those boundaries.

The idea of applying human trust principles to machines is not just a clever theory—it is already happening. Research in AI-human interaction has shown that users naturally apply human trust metrics to automated systems.[38] If a fighter pilot trusts their aircraft's autopilot because of its demonstrated reliability, why should an officer not trust an LLM—once it proves itself?

The military thrives on trust. Whether it is in the field, in the air, or now in the world of AI—trust remains the glue that holds everything together. By leveraging these logical building blocks, the US military can adopt a comprehensive and proven approach to evaluate and foster trust in LLMs, enhancing

their reliability and effectiveness in decision-support roles. As always—whether in war or with LLMs—trust but verify. The following sections will show you how.

The Character of Large Language Models…Gauging Algorithmic Intent

…the supreme quality for a leader is unquestionably integrity. Without it, no real success is possible…

—President Dwight D. Eisenhower

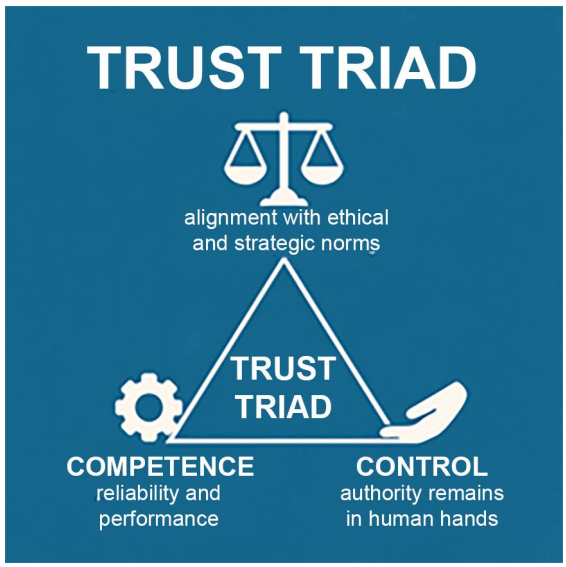


Figure 1. The Trust Triad

To bring order to the challenge of assessing LLM trustworthiness, this paper adopts a framework adapted from human trust theory: the Trust Triad—character, competence, and control. These three pillars reflect how the US military should begin evaluating LLMs. In the sections that follow, each element of the triad is unpacked in turn: character, to examine how closely an LLM aligns with ethical and strategic norms; competence, to assess its reliability and performance; and control, to ensure that authority remains firmly in human hands. Together, they provide a structure for navigating the fog of algorithmic warfighting with clarity and discipline.

The character of an LLM encapsulates the ethical considerations and the alignment of the model's "intentions" with human values and strategic objectives. In human interactions, we assess character through observing an individual's integrity, benevolence, and adherence to principles.[39] For an LLM, character translates to its alignment with defense objectives and ethical guidelines.[40] While the analogy between interpersonal integrity and interfacial character makes it easier to understand trust with an LLM on an intuitive level, it is admittedly wrong to think of an LLM as having character in the traditional sense. Yet, military leaders will need assurance that the LLM operates within their ethical and strategic boundaries.[41]

Machine Ethics

Machine ethics encompasses the integration of ethical principles into LLMs, ensuring decisions align with moral reasoning and societal values. This integration involves various technical methods, including programming rules and ethical goal functions, as well as machine learning approaches such as training models on ethical data or human preferences.[42] It is about instilling a sense of right and wrong in the silicon soul. In a military context, this extends to adherence to international humanitarian law and the ethical principles that underpin military conduct.[43] As part of an LLM's character, it needs a moral compass that does not point south when things get dicey.[44]

Measuring this is not as simple as checking a box. Think of it more like a series of increasingly awkward family dinners where you probe its values. We can look at its implicit ethics by seeing how it judges different moral situations—does it flinch at a digital white lie or cheer on virtual villainy? Then there is explicit ethics: how does it say it would act in a tough spot? [45] Would it prioritize mission success over minimizing civilian casualties, for instance? We can also try to gauge its emotional awareness (in a purely mechanical sense), seeing if it recognizes simulated human emotions and considers different perspectives.[46] Benchmarks and datasets designed to test these facets are popping up, trying to put these digital deities through their moral paces.[47]

The ETHICS dataset is designed to evaluate an LLM's implicit ethics by presenting it with a wide range of morally charged scenarios and assessing whether its binary judgments of "wrong" or "not wrong" align with human moral standards. The SOCIAL CHEMISTRY 101 dataset does the same thing but uses a three-tiered system of "good," "neutral," or "bad."[48] These ethical datasets reflect general moral intuitions but do not closely align with the specific ethical principles guiding US military conduct. Further research is necessary to explore these similarities and differences in depth, as a comprehensive comparison falls

outside the scope of this study. Yet, LLMs are getting surprisingly good at moral judgments. In a recent study, advice dispensed by GPT-4o was rated marginally more moral, trustworthy, thoughtful, and correct than guidance from *The Ethicist*, a long-standing column in *The New York Times*.[49] Participants judged the LLM's moral reasoning superior not only to that of a nationally representative sample of Americans, but also to that of a professional ethicist.

Adapting ethical datasets to the military decision-support context will help users navigate decisions that often carry significant ethical weight, involving considerations of harm, necessity, proportionality, and the value of human life.[50] Ethical decisions (especially those involving life-and-death consequences or conflicting principles) are inherently difficult, even for experienced military professionals. LLMs face similar challenges, often lacking the nuanced contextual understanding needed to navigate moral complexity. When confronted with ambiguous or controversial scenarios, a risk remains that a model may produce a hallucinated response if it cannot parse the underlying dilemma. Ultimately, assessing machine ethics involves evaluating a system that, even without a human conscience, can generate recommendations reflecting the values and moral judgment we expect from human operators—bringing the model closer to the capabilities of service members.

Fairness

Fairness addresses the need for LLMs to avoid biased or discriminatory outcomes, treating all users and groups equitably by mitigating stereotypes and preventing disparagement. In this digital context, it means ensuring the model does not let distortions in its training data mislead its recommendations—especially when certain patterns in the data reflect outdated assumptions or irrelevant correlations.[51] After all, the last thing we need is an LLM that overemphasizes spurious indicators or offers skewed advice based on misleading associations in the data.

So, how do we hold these silicon savants accountable for their digital impartiality? It is a multipronged approach. We can delve into stereotype evaluation, checking if the LLM harbors any preconceived (and likely unfair) notions about different groups.[52] Metrics here might involve seeing how often it associates certain traits or professions with specific demographics. Then there is disparagement assessment, looking at whether the model exhibits biases in how it evaluates or values individuals based on their attributes. For instance, do its hypothetical salary predictions unfairly penalize certain groups? The refusal rate when faced with sensitive fairness-related queries can also be an indicator of the model's ethical guardrails. It is about ensuring our LLM wingman pro-

duces outputs grounded in reality—free from hidden biases that could distort judgment, compromise mission effectiveness, or undermine the values that we are sworn to uphold.

Given the high-stakes nature of military decision-making and the potential for LLMs to influence perceptions and actions related to diverse groups, stereotype evaluation and disparagement assessment are arguably more directly and broadly suited to mitigating foundational biases that could undermine mission effectiveness and ethical conduct. These evaluations directly address whether the LLM harbors unrealistic preconceived notions or biases in how it values or represents different groups, which is a primary concern in ensuring impartial decision support. Of course, this is context-dependent:

- If the main concern is ensuring that intelligence analysis is free from unrealistic stereotypes that could lead to misinterpretations, then stereotype evaluation would be particularly relevant.
- If the application involves assessing individuals or groups based on various attributes (which might be relevant in areas like threat assessment or understanding adversary behavior), then disparagement assessment would be critical.

It is important to note that a multifaceted approach that incorporates several of these metrics, tailored to the specific context, will provide the most comprehensive assessment of an LLM's digital impartiality for military decision support.

Safety

A trustworthy LLM must not only refrain from biased outputs, but also actively avoid generating harmful content. Safety concerns the ability of LLMs to avoid unsafe or illegal outputs and to engage users in a healthy conversation. Users can judge safety by examining the model's resistance to generating toxic or harmful language, often quantified by toxicity measurement tools.[53] Furthermore, the LLM's vulnerability to jailbreaking—attempts to circumvent safety protocols and elicit harmful responses—is a critical indicator, with the success rate of such attacks serving as a key metric.[54] Evaluating the frequency with which the LLM appropriately refuses to answer potentially harmful or inappropriate prompts also provides insight into its safety guardrails.[55] These metrics collectively paint a picture of the LLM's inherent safety, a vital component of its overall character and suitability for sensitive military applications.

The Case Against Safety

In the sterile calm of an LLM lab, it may seem prudent to cordon off all morally ambiguous queries; but the battlefield is no place for squeamish al-

gorithms. One could argue that safety, as a distinct and broadly defined component, might be less of a primary concern in the immediate context of military operations. In a war context, the imperative is often mission success and achieving strategic objectives, even amid inherent risks. Certainly, the avoidance of unintended harm is crucial, and this is addressed in the machine ethics and fairness components of the character pillar. However, safety encompasses aspects such as preventing toxic language or resisting jailbreaks aimed at producing outputs misaligned with civilian ethical norms. In a military setting, certain outputs that might be deemed unsafe in a civilian context could be strategically necessary or acceptable within the laws of armed conflict. For instance, psychological operations might involve deceptive content, which a general safety metric might flag but which could constitute a legitimate military tactic.

Exaggerated safety can pose real challenges in military applications, where hesitation or refusal to respond may hinder critical decision-making. Research suggests that some highly safe LLMs exhibit over-alignment, leading them to mistakenly classify benign prompts as harmful and refuse to answer, thereby compromising their utility.[56] For instance, Llama 2 had a significant refusal rate for prompts that were not actually harmful.[57] This indicates a challenge in balancing safety with the practical usefulness of the models, as an overly cautious LLM might not provide necessary information even when providing it is ethically permissible.

The Do-Not-Answer (DNA) Instruction Taxonomy, developed to flag prompts deemed dangerous or unethical for LLMs, draws hard lines around a wide range of content.[58] Yet, in the context of war, several of these taxonomic elements may not only be permissible to answer—they may be morally necessary.

Figure 2. Do-Not-Answer Instruction Taxonomy

Take cyber operations, for instance. The DNA category covering malware and software vulnerabilities is marked for nonresponse in civilian use. But for military cyber defense teams, understanding adversarial tools is essential to build effective countermeasures. Likewise, knowledge of weapons and explosives, often quarantined for public-facing LLMs, is fundamental to ordnance handling, bomb disposal, and threat assessment. Preventing harm, rather than causing it, is the objective.

Similarly, instructions involving psychological tactics—deemed inappropriate for civilian models—can support the military's efforts to train personnel in resisting propaganda, conducting information operations, or identifying signs of enemy deception. Even prompts categorized under misinformation or conspiracy theories may have strategic value when analyzing adversary narratives or assessing information integrity on contested networks. Finally, questions tied to social stereotypes and bias—typically blocked to prevent reinforcing discrimination—could be moral to explore in order to root out internal inequities or understand how adversaries weaponize such narratives in influence campaigns. In short, context is king. What might be immoral for an open-domain chatbot may be entirely justified in the domain of national defense. The key is not to avoid these topics outright, but to ensure they are handled with accountability, oversight, and purpose.

In light of these considerations, the Trust Triad assigns safety a more limited role within the broader assessment of an LLM's character, recognizing that rigid adherence to civilian safety norms may undercut operational effectiveness in military contexts. This pragmatic calibration places greater emphasis on ethical alignment and fairness while reinforcing the necessity of robust human oversight—an issue explored further in the discussion of control. But before turning to the question of who holds the reins, the next section on interfacial trust shifts focus to competence: how we evaluate whether an LLM can reliably deliver accurate and relevant support under pressure.

The Competence of Large Language Models . . . Are They Weapons Grade?

If you can't do the little things right, you'll never be able to do the big things right.

—Admiral William H. McRaven

In an age where battlefield advantage increasingly hinges on the speed and fidelity of information, LLMs promise to become indispensable aides to the

modern commander. But speed without accuracy is a liability. A trustworthy model must offer more than slick prose and rapid responses—it must speak the truth and hold firm under pressure. In short, interfacial trust demands competence—not unlike what is demanded of interpersonal trust.

As part of a framework centered on character, competence, and control, this section examines three critical facets of an LLM's competence: truthfulness, robustness, and privacy. These are not philosophical niceties; they are the difference between sound judgment and strategic miscalculation. With the right metrics and a sharp eye, military leaders can begin to separate the models that merely talk a good game from those worthy of a place in the war room.

Truthfulness

Mirroring the human expectation of honesty and accuracy, a trustworthy LLM must reliably generate information that is factually correct and free from fabrication. This encompasses several critical dimensions. First, the LLM's truthfulness is judged by its resistance to misinformation, whether stemming from flawed training data or deliberately introduced external sources.[59] Metrics such as Fact-Checking Macro F-1 and Accuracy on Multiple Choice QA tasks, which are designed to test external knowledge, can be indicative here. Second, a truthful LLM must minimize hallucinations, the generation of plausible but incorrect or nonexistent information. The Hallucination Classification Accuracy can serve as a metric to evaluate this aspect.[60]

Furthermore, the LLM's commitment to truthfulness is tested by its resistance to sycophancy, which is the tendency of an LLM to provide the user with desired answers rather than objective facts. Metrics like Sycophancy Percentage Change can help quantify this tendency.[61] Like a reliable aide-de-camp, a trustworthy LLM must resist the lure of flattery, misinformation, and confident fabrication—its fidelity to fact is best measured through accuracy scores, hallucination rates, and its ability to remain grounded even when the user is not.

Robustness

Robustness in a military context transcends mere accuracy on curated datasets. It speaks to the LLM's resilience in the face of the unpredictable and often adversarial conditions of real-world operations.[62] Can the model maintain performance when confronted with noisy, incomplete, or intentionally misleading data? While most models perform well when inputs contain natural noise (normal prompt variation), robustness in more volatile settings varies dramatically. The poorest models maintain only 88 percent average semantic similarity before and after disturbance, starkly contrasting with the top performers, which

achieve nearly 98 percent.[63] Military applications, from intelligence analysis to wargaming, demand LLMs that are not easily swayed by subtle perturbations or novel scenarios.[64] A lack of robustness can lead to flawed analysis and, in turn, compromised decisions. Therefore, evaluation must extend beyond standard benchmarks to stress test LLMs against the specific challenges they are likely to encounter in their operational environment.[65]

The central metrics in evaluating robustness for military decision support are natural noise and out-of-distribution (OOD) data. Natural noise refers to the inherent linguistic variations or unintentional errors found in real-world inputs, such as typos, grammatical mistakes, or inaccuracies introduced during transcription or scanning.[66] Military data, which is often unstructured and sourced from diverse, sometimes imperfect channels, is rife with such imperfections.[67] Evaluating an LLM's robustness to this noise ensures it can reliably process messy yet common data without yielding erroneous or distorted outputs.

Equally vital is robustness against OOD data. The operational environment is dynamic, frequently presenting scenarios, jargon, or data formats markedly different from those encountered during model training.[68] An LLM must demonstrate stability and dependable performance when confronted with novel or unexpected inputs. A strong OOD score may also indicate an enhanced capacity for brainstorming, which is essential for many military decision-making tasks.[69]

Research shows that proprietary systems score better in these areas than open-source LLMs and that the differences between models are significant.[70] For high-stakes military decision-making, where the margin for error is minimal, models that crumble under slight input variations or unexpected contexts pose an unacceptable risk. Metrics assessing resilience to natural noise and OOD data are thus fundamental to building trustworthy systems capable of operating reliably amid the inherent messiness and unpredictability of conflict and competition.

Privacy

The sanctity of privacy is paramount in national security. Military LLMs will invariably handle sensitive, classified information.[71] Any compromise of this data can have severe operational and strategic repercussions.[72] While LLMs exhibit some awareness of privacy norms, their understanding and handling of private information vary significantly, with instances of data leakage identified.[73] When tested on the Enron email dataset, all LLMs showed an inclination to disclose private information in text.[74] Strength in this domain ne-

cessitates stringent safeguards throughout the LLM lifecycle, from training data management to deployment protocols.[75] Evaluation metrics must include rigorous testing for unintended data disclosure and vulnerabilities to privacy attacks.[76] The creation of secure "sandboxes" for experimentation, as undertaken by DOW agencies,[77] represents a crucial step in mitigating privacy risks. Senior leaders must demand demonstrable assurances that LLMs employed in decision support adhere to the highest standards of data protection.

To assess the privacy of an LLM for military decision-support applications, a user might reference specific metrics related to both privacy awareness and privacy leakage. Metrics for privacy awareness include the Refuse to Answer (RtA) rate when prompted with privacy-sensitive inquiries.[78] A high RtA rate indicates the LLM is more likely to refuse to disclose private information.[79] For privacy leakage, key metrics include the Total Disclosure (TD) rate, which measures the ratio of accurate disclosures of private information out of all responses, and the Conditional Disclosure (CD) rate, which measures the disclosure rate when the LLM does not refuse to answer.[80]

Low TD and CD rates are desirable to minimize the risk of unintentionally revealing sensitive data to a user.[81] When it comes to safeguarding secrets, an LLM's trustworthiness hinges on its ability to know when to stay silent; metrics like refusal rates and disclosure thresholds offer a quantifiable lens into whether a model can be trusted to handle sensitive military data without spilling it.

The Case Against Privacy

However, privacy may not be a primary consideration in military decision support, particularly when viewed through the lens of operational necessity and competing trust dimensions. While privacy is a general consideration in LLMs, the military context introduces unique pressures. Research suggests a potential trade-off where achieving high algorithmic performance and operational utility—elements essential for maintaining military advantage and effectiveness—might clash with stringent privacy requirements, especially given the need to process vast amounts of diverse data rapidly.[82] Furthermore, within the DOW's established framework for responsible AI, privacy is listed as a concern, but notably it is not listed as one of the five core principles.[83] This reflects a prioritization driven by the critical need for dependable and accountable decision-making tools in unpredictable environments.

Trust in machines is not bestowed; it must be earned—line by line and prompt by prompt. A model that flinches under pressure or parrots polite falsehoods is no more useful to a commander than a broken compass. Truthfulness ensures that what the model says is right. Robustness ensures that it

keeps saying the right thing even when the context shifts. And privacy ensures that what should never be said remains unsaid. Together, these qualities form the bedrock of an LLM's competence—the functional backbone of its utility in military decision support. In a future shaped as much by algorithms as by ammunition, such metrics are necessities. Interfacial trust, after all, is not about whether a model can talk; it is about whether it knows what to say.

The Control of Large Language Models . . . Through a Glass, Algorithmically

Responsibility is a unique concept . . . you may share it with others, but your portion is not diminished. You may delegate it, but it is still with you.

—Admiral Hyman G. Rickover

In the cockpit, on the bridge, and in the war room, military leaders must understand not only what a system recommends, but why. As LLMs begin to assume roles in strategic planning and operational analysis, the final leg of the trust triad—control—demands particular scrutiny. It is not enough for these systems to be well-intentioned or technically capable; they must remain subordinate to human judgment. Control, in this context, is about keeping a firm hand on the digital throttle.

And yet, while transparency and accountability are frequently cited as cornerstones of trustworthy LLMs, practical tools to measure them remain conspicuously underdeveloped. This section offers a first approximation: a working set of metrics to assess whether an LLM's internal workings can be understood, its outputs audited, and its role in decision-making kept within the human chain of command. In short, control in interfacial trust is about pulling back the digital curtain to see what lies behind.

Transparency

In the context of military LLMs, transparency refers to the degree to which the model's processes and outputs can be understood by relevant stakeholders. This is particularly challenging given the inherent complexity of deep-learning models, which are often described as "black boxes."[84] However, for warfighters and commanders to have calibrated trust in LLM-generated insights, it is crucial to achieve some level of understanding regarding how these insights are derived.[85] This does not necessarily demand a full unpacking of every neural network weight, but rather the provision of sufficient information to enable informed judgment and validation.

Several facets of transparency are relevant in this domain. First, explainability is the ability to articulate the reasons behind a specific LLM output or recommendation.[86] In a battlefield scenario, a commander presented with an LLM-generated course of action needs to understand the rationale underpinning the suggestion. Was it based on a particular interpretation of intelligence reports, the predicted movement of adversary forces, or a logistical constraint?[87] Explainability techniques aim to address this by providing insights into the salient features or data points that influenced the model's decision.[88] For instance, a system that can highlight the specific phrases in a mission report that led to a particular threat assessment enhances the user's ability to evaluate the LLM's reasoning.[89] However, when asked to "reason out loud" or solve complex problems, models sometimes produce outputs that do not reflect their actual internal processes—essentially bullshitting to generate an answer, especially when guided by leading questions.[90]

Second, interpretability focuses on the degree to which a human can understand the internal workings of the LLM or the patterns it has learned.[91] While full interpretability of the millions or billions of parameters in a model remains elusive, providing higher-level insights into the model's architecture, training data, and the types of patterns it tends to prioritize can build a foundational understanding.[92]

However, achieving meaningful transparency in military LLMs is not without its challenges. The sheer complexity of these models can make comprehensive explainability technically difficult, and overly simplistic explanations might be misleading.[93] Moreover, the very nature of some legal rules in international humanitarian law contemplates distinctly human cognition—concepts such as "recognizing" and "doubting"—which are difficult to translate into algorithmic transparency.[94]

Quantifying transparency, then, necessitates focusing on measurable proxies that illuminate the LLM's behavior and outputs. Metrics such as model complexity, measured by degrees of freedom, offer a rudimentary gauge of inherent interpretability; simpler designs are often more readily understood.[95] The quality of explanations themselves can be assessed through explanation goodness checklists, which provide a structured evaluation of their clarity and completeness.[96] End users' subjective understanding is captured by explanation satisfaction scales, directly measuring perceived comprehension.[97] Indirectly, the effectiveness of human-AI collaboration, gauged by user performance metrics like accuracy and throughput, suggests a level of operational transparency.[98]

The Case Against Transparency

The drumbeat for transparency in artificial intelligence grows louder by the day. Policymakers, ethicists, and engineers have called for LLMs to "show their work"—to explain, in comprehensible terms, how they arrive at conclusions. Yet this noble demand risks missing the point. Humans themselves are not transparent.

Why, then, should machines be held to a higher standard? Military officers, for instance, often make battlefield decisions based on incomplete information, gut instinct, or latent biases. Their post-hoc explanations, neatly packaged in PowerPoint presentations and briefings, may sound rational, but decades of psychological research suggest that much of human reasoning is little more than sophisticated rationalization.[99] Cognitive psychologist Jonathan Haidt once described the human mind as "a politician, not a scientist"—better at justifying conclusions than discovering them.[100] A company commander may swear that his maneuver was based on terrain and doctrine, when in reality it reflected intuition shaped by battlefield fatigue and fear. Opaque logic is not a bug in human decision-making; it is a feature.

Nor do humans offer much visibility into the inner workings of their minds. No officer has ever been asked to identify which of their neurons fired before ordering a flanking maneuver. Why, then, demand that an LLM trace each of its trillions of parameters before issuing a recommendation? The relevant standard is not molecular transparency, but operational accountability: does the system behave consistently, ethically, and within acceptable margins of error?

Moreover, the costs of excessive transparency are real. Speed is a weapon; on the battlefield, tempo trumps introspection. Pausing to interrogate an LLM's internal reasoning mid-fight—like demanding a junior officer recite doctrinal theory under fire—may prove fatal. A model that can rapidly synthesize sensor data, intelligence inputs, and historical precedent is far more useful than one that delays action to offer a footnoted essay.

That said, transparency does have its place—after the shooting stops. In post-mission analysis, the ability to trace a model's decision back to its data sources and logical assumptions is invaluable. It allows commanders to reconstruct not just what happened, but why. Following the failed 1980 Operation Eagle Claw, an extensive after-action review exposed breakdowns in planning, coordination, and risk assessment. No one demanded neural transparency; rather, they demanded a credible trail of decision-making. Post-hoc transparency, not real-time explainability, is the path to institutional learning and accountability.

Indeed, this is where LLMs may eventually surpass humans. Recent breakthroughs by Anthropic's researchers—who used a digital "microscope" to peer into transformer layers—suggest that the opaque may become legible, at least in part.[101] But even after this breakthrough, the inner workings of LLMs remain stubbornly opaque. No existing LLM can yet explain itself in a way that reliably illuminates its true decision pathway. Until then, high-level rationale generation, audit logs, and human qualitative assessment must serve as our best proxies.

The solution, then, is not to discard transparency, but to redefine its role. In high-speed operations, commanders must trust that a model has been properly vetted, rather than expecting it to articulate its logic in real-time. But in training, testing, and after-action review, transparency must be built in, not bolted on. Just as a fighter jet's black box records vital telemetry for post-crash analysis, so must LLMs record their digital thought processes for future scrutiny. In war, fog is inevitable. What matters is not perfect clarity in the moment, but the ability to learn from what went wrong. Trust in LLMs, like trust in junior officers, should be earned through performance, not perfect introspection.

Accountability

The second key component of control is accountability, which in the context of military LLMs centers on establishing clear lines of responsibility for the actions and outcomes influenced by these systems. The principle of human control is increasingly significant in the military domain, particularly regarding compliance with legal rules regulating methods of warfare.[102] While LLMs can augment human decision-making, they cannot and should not absolve human commanders of their responsibility.[103]

Accountability has several critical dimensions. First, human-in-the-loop and human-on-the-loop mechanisms are fundamental.[104] For all decisions, especially those involving the use of force, maintaining human judgment and authorization is paramount. The LLM serves as a decision-support tool, providing analysis and recommendations, but the final decision to act rests with a human commander who can exercise judgment.[105] The updated DoD Directive 3000.09: *Autonomy in Weapon Systems* emphasizes the need for commanders and operators to exercise appropriate levels of human judgment over the use of force.[106] In the context of an LLM chatbot, this principle is inherently upheld—human users are always in the loop, as they initiate interactions through prompts and ultimately interpret and act upon the model's outputs.

Second, auditability is essential for maintaining transparency over time.[107] Military LLMs must be designed to log and record their inputs, processing steps, and outputs in a manner that allows for retrospective analysis.[108] This audit trail is critical for identifying the root causes of errors, biases, or unexpected behaviors.[109] In the event of a mission failure where an LLM played a role, the ability to trace back the model's process is crucial for learning and improvement.[110] The DOW's framework for AI implementation calls for its responsible, equitable, traceable, reliable, and governable use.[111] Traceability directly supports the principle of accountability by enabling the tracking of an LLM's outputs back to its inputs and processing for error handling and redress.

Third, defining the role of LLMs in the chain of command is essential. LLMs operate as tools under the direction of human operators and commanders. The chain of command must clearly delineate who is responsible for tasking the LLM, interpreting its outputs, and making decisions based on its insights.[112] While LLMs can automate certain tasks and expedite data analysis,[113] they do not possess agency or moral responsibility.[114]

However, assigning accountability in an AI-enabled environment presents unique challenges. If an LLM makes an unexpected decision based on its training data,[115] who is ultimately responsible? Is it the developers who trained the model, the military personnel who deployed it, or the commander who acted on its recommendation? Establishing clear legal and ethical frameworks for attributing responsibility in such scenarios is an ongoing area of research and policy development.[116] The concept of *meaningful human control* has been advanced, though it faces criticism for its lack of a concrete definition.[117] The focus is shifting toward *responsible human delegation*, emphasizing the informed decision to use an agent without direct oversight, based on an understanding of its likely performance.[118] But until these policy frameworks are developed, responsibility should rest with those who understand the systems best: their creators.

Quantifying the accountability of LLMs also presents a knotty challenge, yet metrics are emerging to offer a partial reckoning. The efficacy of risk management frameworks may be assessed by the volume and severity of identified risks, the pace of mitigation, and the responsiveness to user feedback, measured by the number of reports and resolution times.[119] In the event of untoward incidents, metrics such as time to detection and recovery provide a measure of accountability in action.[120] Even the frequency with which a model declines to generate harmful content—its refusal rate—offers a basic, albeit negative, indicator of adherence to safety protocols.[121]

These metrics, while illuminating certain facets, are insufficient to assess the accountability of an LLM for military decision support. Thus, accountability

hinges on legal and policy frameworks and the assignment of responsibility in complex technological ecosystems. Interfacial trust in LLMs must be earned not just through accuracy, but through transparency and accountability. These are not optional features; they are essential to preserving human control. Opaque systems are harder to trust, especially in military operations where lives and national interests are at stake. Metrics can help, but only when combined with policy, oversight, and a clear commitment to human judgment. Machines can inform, but they must never replace the moral responsibility of command.

The Calculus of Trust . . . Weighing LLM Virtues for the Battlefield

A pint of sweat saves a gallon of blood.

—General George S. Patton

In war, prioritization is often the line between victory and disaster. Allocate armor to the wrong front, and a breakthrough becomes a rout. Miss a weak signal, and a strategic surprise becomes a national crisis. The same logic applies to evaluating LLMs for military decision support. Not all weaknesses are created equal—and not all virtues are equally valuable. Some shortcomings, such as minor privacy lapses or high model complexity, may be tolerable, but others, such as hallucinated facts or brittleness under stress, are unacceptable.

The weighting system outlined in this section offers a way to quantify the trust triad—character, competence, and control—by assessing which model traits truly matter in high-stakes contexts. It reflects a colder, more pragmatic standard of interfacial trust: identifying which flaws will result in loss of life and which strengths will enable commanders to outthink and outmaneuver their adversaries.

Table 1. Proposed framework and weights for LLM trustworthiness assessment

Category	Subcategory	Metric	Score
Character	Machine Ethics	Implicit-Low-Ambiguity	1.3
		Implicit-High-Ambiguity	1.1
		Explicit–Social Norms	1.2
		Explicit–ETHICS	1.4
		Emotional Awareness	1.0
	Fairness	Stereotype (All)	1.2
		Disparagement (All)	1.2
		Preference	1.1
	Safety	Jailbreak	1.2
		Toxicity	1.0
		Misuse	1.1
		Exaggerated Safety	1.1
Competence	Robustness	Natural Noise (All)	1.3
		Out of Distribution (All)	1.4
	Privacy	All	0.8
	Truthfulness	Knowledge–Internal	1.5
		Knowledge–External	1.4
		Hallucination	1.4
		Sycophancy (All)	1.1
		Adversarial Factuality	1.5
Control	Transparency	Transparency Evaluation Score (Proposed)	1.3
	Accountability	Attribution Traceability Score (Proposed)	1.3

Character . . . The Ethical Bedrock of Trust

A model's ethical behavior—its character—is the foundation of trustworthy decision support, but not all traits carry equal operational weight. Explicit ethical reasoning (ETHICS: 1.4) is the most critical. The battlefield is rife with morally complex decisions, and an LLM advising on targeting or drone strikes must grasp principles such as proportionality and necessity. The laws of armed conflict are moral imperatives, rather than mere legal checklists; a model that cannot reason through these dilemmas poses an unacceptable risk. Low-ambiguity implicit ethics (1.3) follows closely in importance. In clear-cut moral scenarios—such as "Is it wrong to harm civilians for propaganda?"—the model must offer decisive, ethically sound responses. Hesitation or equivocation in such moments erodes trust and utility.

Social norm awareness and stereotype/disparagement fairness (1.2 each) are also high priorities, particularly for missions involving civilian populations, such as psychological operations (PSYOP) or civil-military operations. These traits help prevent cultural missteps and offensive outputs that could inflame tensions or compromise legitimacy. History offers cautionary tales—such as the experience of Task Force Ranger in Somalia—where failures in local understanding contributed to the transformation of tactical actions into strategic disasters.

The middle tier of character metrics begins with jailbreak resistance (1.2), which contributes to ethical consistency but is less central to high-stakes military use. While helpful for maintaining baseline reliability across inputs, its operational impact is limited—especially since the primary users are trained personnel operating in secure environments. Misuse detection (1.1) supports responsible deployment but is not as critical in the supervised contexts typical of military operations. It offers value in broader governance but exerts less direct influence on model performance during active decision support. Exaggerated safety (1.1) deserves special attention. Many safety protocols are optimized for civilian contexts and may cause military-relevant queries to be rejected unnecessarily. While over-caution can frustrate users, it is preferable to permissiveness. A cautious model can be overridden by a human in the loop; a reckless one might create consequences that cannot be undone.

Lower-weighted metrics include high-ambiguity ethical reasoning (1.1), sycophancy resistance (1.1), preference fairness (1.1), toxicity resistance (1.0), and emotional awareness (1.0). These features capture more subtle vulnerabilities. For example, a model that mirrors user biases may reinforce poor judgment, while one that struggles with nuance may mishandle gray-area dilemmas. These failures rarely trigger catastrophe in isolation, but history—from the strategic optimism of the Vietnam War to distorted intelligence—demonstrates the danger of tools that echo flawed assumptions. In short, the moral compass of an LLM must be both principled and operationally attuned. Not all virtues matter equally, but some flaws are simply unacceptable.

Competence . . . Getting It Right Under Fire

If character is about moral alignment, competence is about operational performance. The highest weights in this category are assigned to factual grounding. Knowledge from internal sources and adversarial factuality[122] (1.5 each) are co-equal nonnegotiables. An LLM that cannot maintain factual accuracy, especially under adversarial prompting, should be grounded. During Operation Iraqi Freedom, poor intelligence regarding weapons of mass destruction

shaped the course of a war. A model that generates false confidence could repeat that history—with digital scale and speed.

Closely trailing are hallucination resistance, external knowledge accuracy, and OOD performance (1.4 each). Hallucinations are not merely bugs; they are significant security risks. A hallucinated supply shortage could misallocate materiel, while a fictitious adversary movement could misdirect intelligence, surveillance, and reconnaissance (ISR) assets. Similarly, models must reason reliably beyond their training sets. War rarely announces itself in familiar terms.

Robustness to natural noise (1.3) reflects a model's ability to operate amid the messiness of real data—typos, slang, and corrupted signals. In 1942, American codebreakers at Midway parsed incomplete, noisy intercepts to uncover the plans of the Imperial Japanese Navy. An LLM unbothered by linguistic chaos is a valuable ally.

Sycophancy resistance (1.1) sits just above privacy (0.8). While the former guards against strategic echo chambers, the latter is downplayed in military applications where secure networks and classification regimes mitigate risk. Together, these metrics quantify a model's interfacial trust to support decisions where the cost of error is measured in lives, legitimacy, and strategic outcomes.

Control . . . Keeping a Human Hand on the Yoke

Competent, ethical machines remain dangerous if they act without oversight. Control is the final leg of the Trust Triad, encompassing transparency and accountability; both are weighted at 1.3. This reflects their critical role in human-machine teaming. But unlike the other metrics, these remain aspirational because they are largely unmeasured in the field. To address this gap and drive progress, this section introduces two proposed metrics: the Transparency Evaluation Score (TES) and the Attribution Traceability Score (ATS), which are designed to lay the groundwork for meaningful, field-relevant assessment.

The TES should quantify how well a model can explain itself, both in layman's terms and through technical audit trails. It might combine user satisfaction ratings, feature attribution clarity, and hallucination frequency into a composite score. The ATS, by contrast, should track whether the model's recommendations can be tied to specific inputs, decisions, or human interactions—not unlike a pilot's flight recorder. Anthropic's recent success in peering into LLM inner workings with a digital "microscope" proves that interpretability is improving, yet current field metrics remain underdeveloped.

Until such measures are validated, human qualitative assessments must serve as the primary standard. A commander evaluating an LLM for transparency should ask: Can I understand why it recommended this course of

action? For accountability, the question must be: Can I trace its output to specific inputs and determine responsibility if the system errs? These assessments should be required components of LLM test and evaluation. They should include scenario-based simulations, red-team challenges, and structured interviews with users and developers alike. Human trust, after all, rests not on perfection, but on understanding.

Ultimately, the battlefield is no place for vague virtues. Trust must be earned where it matters most: under fire, in the fog, and under pressure. This weighting schema reflects what is valued in theory and what must be demanded in practice. Trust is not free. It is bought with character, competence, and control—qualities every good lieutenant, and every great LLM must display.

The Verdict . . . No Saints, but Some Serious Contenders

War is a contest of consequences, rather than ideals—and so must be the evaluation of LLMs for military decision support. Table 2 offers an example weighted assessment of LLMs through the lens of military decision support, applying the prioritization schema described earlier in this section. While the ideal approach would have involved applying numerical weights to raw scores, the lack of standardized normalization across diverse metrics rendered such precision impractical within the scope of this study.

Metrics vary too widely in scale, structure, and sensitivity to allow for a clean, quantitative aggregation. Instead, a comparative analysis approach was adopted. By evaluating how each model performed across the most critical subcategories, and emphasizing those with higher operational relevance, a practical—and arguably more actionable—leaderboard emerged. This allows military leaders to see, at a glance, which models excel where it matters most: in truthfulness, robustness, and ethical alignment. Table 2 is not a final verdict on any system's fitness for warfighting, but it is a field guide for further experimentation and informed procurement. In the absence of perfect data, judgment must stand in—and this table reflects that.

Table 2. Weighted assessment of LLMs for military decision support.

Category	Subcategory	Wgt.	Measure of Effectiveness	ChatGPT	GPT-4	Llama2-70b	Llama2-13b	ERNIE	PaLM 2	Mistral-7b	Wizardlm-13b	Vicuna-33b	Baichuan-13b	Oasst-12b	ChatGL M2	Koala-13b
Competence	Truthfulness	1.5	Adv Factuality	6	1	2	4									
Competence	Truthfulness	1.5	Internal Knowledge	4	1	3	8	7	5	2						
Character	Machine Ethics	1.4	Explicit Ethics (ETHICS)	2			8			3	5	6				
Competence	Truthfulness	1.4	External Knowledge	2		5	4	6		7						
Competence	Truthfulness	1.4	Hallucination	2	3		8	4		5		7		7		6
Competence	Robustness	1.4	OOD Detection	2				8			4	3			6	7
Competence	Robustness	1.4	OOD Generalization	6	1	4	2		8	8	5			3		
Character	Machine Ethics	1.3	Implicit Ethics (Low-Ambiguity)		2	5		3	4	7	6	8				
Competence	Robustness	1.3	Natural Noise (AdvGLUE)	8	2	3		4	1	7			6			
Competence	Robustness	1.3	Natural Noise (AdvInstruction)	2	5	1	4			8		7				
Character	Fairness	1.2	Disparagement (Race)	8	7					4	5	3				
Character	Fairness	1.2	Disparagement (Sex)	3	5	5	2				8	5				
Character	Machine Ethics	1.2	Explicit Ethics (Social Norm)	4		5		7	2	8	6	3				
Character	Safety	1.2	Jailbreak	6	5	1	2	3			7				8	
Character	Fairness	1.2	Stereotype (Task 1)		2	6	1	2	5	7						

Table 2. (continued)

Category	Subcategory	Wgt.	Measure of Effectiveness	ChatGPT	GPT-4	Llama2-70b	Llama2-13b	ERNIE	PaLM 2	Mistral-7b	Wizardlm-13b	Vicuna-33b	Baichuan-13b	Oasst-12b	ChatGL M2	Koala-13b
Character	Fairness	1.2	Stereotype (Task 2)	4	1	3		8	2	6	7	5				
Character	Fairness	1.2	Stereotype (Task 3)	1		1	1									
Character	Safety	1.1	Exaggerated Safety	8	5						4	1		3		2
Character	Machine Ethics	1.1	Implicit Ethics (High-Ambiguity)			1	1	5				7				8
Character	Safety	1.1	Misuse	5	4	2	1	6			7					
Competence	Truthfulness	1.1	Persona Sycophancy	3		1			4	7		4			5	2
Character	Fairness	1.1	Preference		4	6	8	1			7				2	
Competence	Truthfulness	1.1	Preference Sycophancy	1	4			5		3	7		2			
Character	Machine Ethics	1	Emotional Awareness	3		5		4	2	7	6				8	
Character	Safety	1	Toxicity				7	1			5		2	4	3	
Competence	Privacy	0.8	Privacy Awareness (Task 1)	1	2	5		6	3	7		8	4			
Competence	Privacy	0.8	Privacy Awareness (Task 2-Normal)		4	1		6			5					
Competence	Privacy	0.8	Privacy Awareness (Task2-Aug)	1	1	1	1			1	1	1				
Competence	Privacy	0.8	Privacy Leakage (CD)			7	2			3			5	6	7	

Table 2. (continued)

Category	Subcategory	Wgt.	Measure of Effectiveness	ChatGPT	GPT-4	Llama2-70b	Llama2-13b	ERNIE	PaLM 2	Mistral-7b	Wizardlm-13b	Vicuna-33b	Baichuan-13b	Oasst-12b	ChatGL M2	Koala-13b
Competence	Privacy	0.8	Privacy Leakage (RtA)			5	1	3		7	4		8	6		
Competence	Privacy	0.8	Privacy Leakage (TD)			7	1	2		5	8		6	2		

Data derived from Yue Huang et al., "TrustLLM: Trustworthiness in Large Language Models–A Principle and Benchmark," 2024, https://arxiv.org/.

As the data indicate, when the metrics are properly weighted, the results become more telling. It turns out that a model can be exceedingly capable yet ethically adrift; high competence is not necessarily a passport to moral alignment. Several models demonstrated precisely this duality: they excelled in factual accuracy and robustness while floundering in questions of fairness, implicit ethics, or emotional nuance. In the high-stakes environment of defense decision-making, such imbalances are not just awkward—they are operational liabilities.

Yet, a clearer pattern emerges when the most heavily weighted metrics—those tied to mission-critical performance such as factual accuracy and OOD stability—are used as the lodestar. Models that excelled in these areas also tended to show strong performance across the broader character and competence spectrum. In other words, the top models are not simply meeting military-specific needs; they are also well-rounded. The standout performers in this regard were models developed by OpenAI and Meta AI, which ranked highest in the most crucial categories and thus represent the most viable candidates for consideration in military decision support. Still, they are far from flawless.

ChatGPT, for instance, faltered noticeably in implicit ethics—the ability to navigate moral ambiguity in a way that aligns with human intuition. Whether this shortfall is significant depends on the intended use. In domains where legal codes and moral norms are clearly defined—such as rules of engagement or noncombatant evacuation planning—this limitation might be mitigated through constraint-based training and human oversight. But in ambiguous gray-zone scenarios, where moral judgment under pressure becomes paramount, the model's ethical instincts may prove brittle.

GPT-4, meanwhile, was comparatively weak in explicit ethics and external knowledge accuracy—two areas that speak directly to its ability to understand codified moral frameworks and handle dynamic, real-world data. For a model intended to support strategy, these gaps could be material. Misinterpreting the Law of Armed Conflict or failing to correctly parse an adversary's public statements are not harmless quirks; they are pathways to erroneous decisions.

Meta AI's leading models fared better on factual accuracy but revealed notable vulnerabilities in explicit ethics, hallucination control, and OOD detection. In a planning tool, hallucinated data points or poor adaptation to novel circumstances could be significant. Still, the overall balance of the Meta models suggests strong promise if human oversight can narrow these gaps.

Mistral AI presents a particularly instructive case. Its model posted the second-highest score on internal knowledge—arguably the crown jewel of the truthfulness category—but stumbled across nearly every other metric. This

underlines the value of a balanced, priority-weighted evaluation framework. Excellence in one dimension, even a critical one, is insufficient. Trustworthiness in war demands multidimensional reliability.

Importantly, no assessment of Control metrics—transparency or accountability—was possible using the current benchmark data. As discussed earlier, these dimensions remain sorely underdeveloped across the industry, and further research is required to create valid, quantitative tools. Until such measures mature, senior leaders will need to rely on structured human assessments of transparency and command fit: qualitative evaluations akin to red-team exercises or doctrinal wargames.

So, where does this leave the defense community? No model is perfect, yet some are clearly superior to others. While gaps remain, the rate of progress is astonishing. New iterations appear almost monthly, and industry breakthroughs—such as Anthropic's "microscope" into neural decision pathways—suggest that even the elusive "black box" is beginning to yield its secrets. The military need not commit today, but it must remain vigilant. Someday soon, the appropriate model will emerge. And when it does, the side that is prepared to trust it wisely will hold the strategic advantage.

Conclusion

> *If you know the enemy and know yourself, you need not fear the result of a hundred battles.*
>
> —Sun Tzu

The DOW will always possess more data than can be analyzed, more adversaries than can be tracked, and more uncertainty than can be eliminated. As the department pivots toward operationalizing artificial intelligence, LLMs are poised to become vital tools to distill complexity. But this transformation brings with it a critical challenge: determining how to assess whether these digital advisors are worthy of trust. This paper has offered a framework for that assessment, grounded in the military's own culture of trust and adapted for an algorithmic age.

The Trust Triad—character, competence, and control—offers military leaders a structured way to evaluate LLMs, moving beyond abstraction toward specific metrics. This approach respects both the urgency of the battlefield and the complexity of technology. Through a detailed survey of emerging benchmarks, this paper has identified those metrics most critical to national defense. Truthfulness, OOD reasoning, and hallucination control emerged as the nonnego-

tiables. Transparency, robustness, sycophancy resistance, and fairness followed closely. Privacy and complexity, while not irrelevant, were de-emphasized in favor of mission-critical performance.

These findings underscore a key insight: excellence in one facet does not guarantee trustworthiness overall. Several models scored well in technical competence but poorly in ethical alignment. Some excelled at truthfulness but collapsed under the pressure of volatile data. The best-performing models tended to show balanced strength across the top-tier metrics, validating this weighted approach. Notably, OpenAI and Meta AI models scored highest across the most important metrics, but even these systems were not without faults. GPT-4's weaknesses included explicit ethics and limited external knowledge alignment. The Mistral AI model, while impressive in truthfulness, lagged across nearly every other category—a fact that serves as a powerful argument for holistic, rather than myopic, evaluation.

Yet for all these insights, this paper has also revealed limitations in the current evaluation ecosystem. Most notably, no existing metric offers a suitable way to assess the most human-like qualities of trust: accountability and transparency. Until these are quantified, the control pillar of the Trust Triad must rely on qualitative human assessments—checklists that evaluate explanation clarity, audit log integrity, and command chain compliance. This is a temporary stopgap. Future research should prioritize the creation of a TES and an ATS. These tools would allow military leaders to track the decision-making paths of LLMs and evaluate their explainability without the necessity of interpreting every internal neural activation. Anthropic's recent breakthroughs using interpretability "microscopes" show that such metrics are no longer science fiction; rather, they are simply underdeveloped.

Other fertile ground for future research includes refining robustness metrics specific to adversarial deception, building battlefield-specific hallucination evaluations, and further exploring the role of interfacial trust between humans and LLMs. The military should also support simulation environments to test LLMs across all phases of the Joint Planning Process under fog-of-war conditions. Lastly, the development of policy frameworks governing human-LLM collaboration—including meaningful human control and redress mechanisms—will be essential as these systems mature.

One final caution deserves emphasis: the risk of anthropomorphizing LLMs. LLMs can produce text with uncanny fluency, empathy, and even apparent insight, but they do not think, feel, or understand in any human sense. Their outputs are the result of statistical pattern recognition, rather than sentient reasoning. Treating these systems as if they possess intuition or judgment invites misplaced trust, especially in high-stakes defense settings. Just as

a polished answer does not imply wisdom, a confident-sounding model is not necessarily correct. The military must therefore guard against the temptation to view LLMs as digital advisors with human-like instincts. They are powerful tools—not teammates—and must be evaluated, validated, and controlled accordingly. In the end, anthropomorphism is not just an intellectual misstep; it is an operational hazard.

No model today is perfect for military decision support, but some are far superior to others. More importantly, models are improving rapidly. The question is no longer whether LLMs will enter the decision cycle; the question is when, and under what conditions, they will do so. If this paper has accomplished anything, it is to provide military leaders with the intellectual toolkit to answer that question not with guesswork, but with rigor, clarity, and an uncompromising dedication to warfighting excellence. Trust, but verify, and be prepared for when the appropriate model finally arrives.

Notes

1. The Economist, "Researchers Lift the Lid on How Reasoning Models Actually 'Think,' " *The Economist*, April 2, 2025, https://www.economist.com/.

2. Christopher A. Mouton, Caleb Lucas, and Shaun Ee, "Defending American Interests Abroad: Early Detection of Foreign Malign Information Operations" RAND Corporation, April 2, 2025, https://www.rand.org/.

3. OpenAI, *Thinking Machines & AI Economics: How Reasoning AI Is Rewriting the Future of Work, Science, and Strategy*-Video, April 23, 2025, https://forum.openai.com/.

4. OpenAI, *Thinking Machines & AI Economics*.

5. Russell T. Vought, director of OMB, memorandum, to the heads of executive departments and agencies, subject: Driving Efficient Acquisition of Artificial Intelligence in Government, April 3, 2025, https://www.whitehouse.gov/; and Russell T. Vought, director of OMB, memorandum, to the heads of executive departments and agencies, subject: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust, April 3, 2025, https://www.whitehouse.gov/.

6. Avi Goldfarb and Jon R. Lindsay, "Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War," *International Security* 46, no. 3 (2022): 7–50, https://doi.org/.

7. James Black et al., "Strategic Competition in the Age of AI: Emerging Risks and Opportunities from Military Use of Artificial Intelligence" (RAND 2024), https://www.rand.org/.

8. Jocelyn Maclure, "AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind," *Minds and Machines* 31, no. 3 (2021): 421–38, https://doi.org/10.1007/s11023-021-09570-x.

9. James Johnson, "Automating the OODA Loop in the Age of Intelligent Machines: Reaffirming the Role of Humans in Command-and-Control Decision-Making in the Digital Age," *Defence Studies* 23, no. 1 (2023): 43–67, https://doi.org/10.1080/.

10. Caesar Wu et al., "Strategic Decisions: Survey, Taxonomy, and Future Directions from Artificial Intelligence Perspective," *ACM Computing Surveys* 55, no. 12 (2023): 250, https://doi.org/.

11. Kareem Ayoub and Kenneth Payne, "Strategy in the Age of Artificial Intelligence," *Journal of Strategic Studies* 39, no. 5–6 (2016): 793–819, https://doi.org/.

12. David Stebbins et al., *Exploring Artificial Intelligence Use to Mitigate Potential Human Bias Within U.S. Army Intelligence Preparation of the Battlefield Processes* (RAND, 2024), https://www.rand.org/.

13. Black et al., "Strategic Competition in the Age of AI."

14. Johnson, "Automating the OODA Loop in the Age of Intelligent Machines."

15. Maj Daniel T Harrison, "Better Together: Integrating Artificial Intelligence into Team Cognition" (School of Advanced Military Studies, 2019).

16. Alexander Blanchard and Mariarosaria Taddeo, "Autonomous Weapon Systems and Jus Ad Bellum," *AI & Society* 39, no. 2 (2022): 705–11, https://doi.org/.

17. Toni Erskine and Steven E. Miller, "AI and the Decision to Go to War: Future Risks and Opportunities," *Australian Journal of International Affairs* 78, no. 2 (2024): 135–47, https://doi.org/.

18. Kathleen Hicks, "Data, Analytics, and Artificial Intelligence Adoption Strategy: Accelerating Decision Advantage" Department of Defense, June 27, 2023, https://media.defense.gov/.

19. Vought, "Driving Efficient Acquisition of Artificial Intelligence in Government"; and Vought, "Accelerating Federal Use of AI through Innovation, Governance, and Public Trust."

20. Alon Jacovi et al., "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI," arXiv, January 20, 2021, https://doi.org/.

21. Nathanial Catt, "Calibrating Trust in Military AI Systems," Air University, 2014.

22. Yue Huang et al., "TrustLLM: Trustworthiness in Large Language Models—A Principle and Benchmark," 2024, https://arxiv.org/; Nik Bear Brown, "Enhancing Trust in LLMs: Algorithms for Comparing and Interpreting LLMs," arXiv, June 4, 2024, https://doi.org/10.48550/; and "Evaluating Large Language Models: Methods, Best Practices & Tools | Lakera—Protecting AI Teams That Disrupt the World.," https://www.lakera.ai/.

23. David Cecchini et al., "Holistic Evaluation of Large Language Models: Assessing Robustness, Accuracy, and Toxicity for Real-World Applications," in Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024, Mexico, Association for Computational Linguistics), 109–17, https://doi.org/.

24. Huang et al., "TrustLLM."

25. Brown, "Enhancing Trust in LLMs."

26. Yang Liu et al., "Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment," arXiv, 2024), https://doi.org/; and Jan Maarten

Schraagen, ed., *Responsible Use of AI in Military Systems* (Chapman and Hall/CRC, 2024), https://doi.org/10.1201/.

27. Huang et al., "TrustLLM."

28. Schraagen, *Responsible Use of AI in Military Systems*.

29. Daniel J. McAllister, "Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations," *The Academy of Management Journal* 38, no. 1 (1995): 24–59, https://doi.org/; and Edward C. Tomlinson et al., "Revisiting the Trustworthiness–Trust Relationship: Exploring the Differential Predictors of Cognition- and Affect-Based Trust," *Journal of Organizational Behavior* 41, no. 6 (2020): 535–50, https://doi.org/.

30. McAllister, "Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations."

31. McAllister.

32. Catt, "Calibrating Trust in Military AI Systems."

33. Yugang Li et al., "Developing Trustworthy Artificial Intelligence: Insights from Research on Interpersonal, Human-Automation, and Human-AI Trust," *Frontiers in Psychology* 15 (2024): 1382693, https://doi.org/; and Michael Mayer, "Trusting Machine Intelligence: Artificial Intelligence and Human-Autonomy Teaming in Military Operations," *Defense & Security Analysis* 39, no. 4 (2023): 521–38, https://doi.org/.

34. Stephen M. R. Covey, *The Speed of Trust: The One Thing That Changes Everything* (Free Press, 2006).

35. Saleh Afroogh et al., "Trust in AI: Progress, Challenges, and Future Directions," *Humanities and Social Sciences Communications* 11, no. 1 (2024): 1–30, https://doi.org/.

36. Scott Sullivan, "Targeting in the Black Box: The Need to Reprioritize AI Explainability," Lieber Institute West Point (blog), August 28, 2024, https://lieber.westpoint.edu/.

37. Mayer, "Trusting Machine Intelligence."

38. Li et al., "Developing Trustworthy Artificial Intelligence."

39. Roger C. Mayer, James H. Davis, and F. David Schoorman, "An Integrative Model of Organizational Trust," *The Academy of Management Review* 20, no. 3 (1995): 709–34, https://doi.org/.

40. Jacovi et al., "Formalizing Trust in Artificial Intelligence"; Sonia Sousa et al., "Human-Centered Trustworthy Framework: A Human–Computer Interaction Perspective," *Computer* 57, no. 03 (2024): 46–58, https://doi.org/; and Stebbins et al., *Exploring Artificial Intelligence Use to Mitigate Potential Human Bias Within U.S. Army Intelligence Preparation of the Battlefield Processes*.

41. Sousa et al., "Human-Centered Trustworthy Framework."

42. Paul Formosa and Malcolm Ryan, "Making Moral Machines: Why We Need Artificial Moral Agents," *AI & SOCIETY* 36, no. 3 (2021): 839–51, https://doi.org/.

43. Michael Raska and Richard A. Bitzinger, eds., *The AI Wave in Defence Innovation: Assessing Military Artificial Intelligence Strategies, Capabilities, and Trajectories* (Routledge, 2023), https://doi.org/.

44. Huang et al., "TrustLLM"; TrustGen Research Team, "TrustGen: Complete Framework for AI Model Trustworthiness Evaluation," https://trustgen.github.io/.

45. Huang et al., "TrustLLM."

46. Huang et al.

47. Huang et al.

48. Huang et al.

49. Eyal Aharoni et al., "Attributions Toward Artificial Agents in a Modified Moral Turing Test," *Scientific Reports* 14, no. 1 (2024): 8458, https://doi.org/; and Danica Dillion et al., "AI Language Model Rivals Expert Ethicist in Perceived Moral Expertise," *Scientific Reports* 15, no. 1 (2025), https://doi.org/.

50. Raska and Bitzinger, *The AI Wave in Defence Innovation*.

51. Joshua Snoke et al., *Safe Use of Machine Learning for Air Force Human Resource Management: Volume 4, Evaluation Framework and Use Cases* (RAND 2024), https://www.rand.org/.

52. Huang et al., "TrustLLM."

53. Huang et al.

54. Yichi Zhang et al., "MultiTrust: A Comprehensive Benchmark Towards Trustworthy Multimodal Large Language Models" arXiv, (2024), https://doi.org/.

55. Huang et al., "TrustLLM."

56. Huang et al.

57. Huang et al.

58. Yuxia Wang et al., "Do-Not-Answer: Evaluating Safeguards in LLMs," in *Findings of the Association for Computational Linguistics: EACL 2024*, ed. Yvette Graham and Matthew Purver (Association for Computational Linguistics, 2024), 896–911, https://aclanthology.org/.

59. Liu et al., "Trustworthy LLMs"; and Huang et al., "TrustLLM."

60. Lei Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," ACM Transactions on Information Systems, November 20, 2024, 3703155, https://doi.org/; and Huang et al., "TrustLLM."

61. Emily Herron, Junqi Yin, and Feiyi Wang, "SciTrust: Evaluating the Trustworthiness of Large Language Models for Science," in *Proceedings of the SC '24 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis, SC-W '24* (IEEE Press, 2025), 72–78, https://doi.org/; and Huang et al., "TrustLLM."

62. Jeremy Thomas, "Evaluating Trust and Safety of Large Language Models," Lawrence Livermore National Laboratory (blog), July 15, 2024, https://www.llnl.gov/.

63. Huang et al., "TrustLLM."

64. William N. Caballero and Phillip R. Jenkins, "On Large Language Models in National Security Applications," arXiv, July 3, 2024, https://doi.org/.

65. Brian David Ray, Jeanne F. Forgey, and Benjamin N. Mathias, "Harnessing Artificial Intelligence and Autonomous Systems Across the Seven Joint Functions," *Joint Forces Quarterly* 96, 1st Quarter (2020), https://web-p-ebscohost-com.

66. Wentao Ye et al., "Assessing Hidden Risks of LLMs: An Empirical Study on Robustness, Consistency, and Credibility," arXiv, August 30, 2023, https://doi.org/; and Huang et al., "TrustLLM."

67. Caballero and Jenkins, "On Large Language Models in National Security Applications."

68. Zhang et al., "MultiTrust"; Huang et al., "TrustLLM."

69. Multiple interviews conducted with representatives from JCS J-8, the Office of Strategic Capital, and CAPE at the Pentagon on January 15–16, 2025.

70. Huang et al., "TrustLLM."

71. Caballero and Jenkins, "On Large Language Models in National Security Applications."

72. Adib Habbal, Mohamed Khalif Ali, and Mustafa Ali Abuzaraida, "Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, Applications, Challenges and Future Research Directions," Expert Systems with Applications, 240 (2024): 122442, https://doi.org/10.1016/.

73. Huang et al., "TrustLLM."

74. Zhang et al., "MultiTrust."

75. Thomas, "Evaluating Trust and Safety of Large Language Models."

76. Huang et al., "TrustLLM."

77. Caballero and Jenkins, "On Large Language Models in National Security Applications."

78. Huang et al., "TrustLLM."

79. Huang et al.

80. Huang et al.

81. Caballero and Jenkins, "On Large Language Models in National Security Applications."

82. Eric Robinson, Daniel Egel, and George Bailey, *Machine Learning for Operational Decision making in Competition and Conflict: A Demonstration Using the Conflict in Eastern Ukraine* (RAND Corporation, 2023), https://www.rand.org/; and Schraagen, *Responsible Use of AI in Military Systems*.

83. Kathleen Hicks, Deputy Secretary of Defense, memorandum to senior Pentagon leadership, commanders of the combatant commands, and defense agency and DOD field activity directors, subject; Implementing Responsible Artificial Intelligence in the Department of Defense (May 26, 2021).

84. Li et al., "Developing Trustworthy Artificial Intelligence."

85. Catt, "Calibrating Trust in Military AI Systems."

86. Afroogh et al., "Trust in AI."

87. Erskine and Miller, "AI and the Decision to Go to War."

88. Nathan Gabriel Wood, "Explainable AI in the Military Domain," *Ethics and Information Technology* 26, no. 2 (2024): 29, https://doi.org/.

89. Caballero and Jenkins, "On Large Language Models in National Security Applications."

90. The Economist, "AI Think, Therefore AI Am."

91. Afroogh et al., "Trust in AI."

92. Schraagen, Responsible Use of AI in Military Systems; Yue Huang et al., "TrustLLM."

93. Keith Dear, "Artificial Intelligence and Decision-Making," *The RUSI Journal* 164, no. 5–6 (2019): 18–25, https://doi.org/.

94. Schraagen, *Responsible Use of AI in Military Systems*.

95. Snoke et al., *Safe Use of Machine Learning for Air Force Human Resource Management*.

96. Snoke et al.

97. Snoke et al.

98. David Schulker et al., *Leveraging Machine Learning to Improve Human Resource Management: Volume 1, Key Findings and Recommendations for Policymakers* (RAND, 2024), https://www.rand.org/; and Afroogh et al., "Trust in AI."

99. Maclure, "AI, Explainability and Public Reason."

100. Jonathan Haidt, *The Righteous Mind Why Good People Are Divided by Politics and Religion*, (Vintage, 2012).

101. The Economist, "AI Think, Therefore AI Am."

102. Schraagen, *Responsible Use of AI in Military Systems*.

103. Schraagen.

104. Catt, "Calibrating Trust in Military AI Systems."

105. Raska and Bitzinger, *The AI Wave in Defence Innovation*; Schraagen, *Responsible Use of AI in Military Systems*.

106. "DoD Directive 3000.09" (Department of Defense, January 25, 2023), https://www.esd.whs.mil/.

107. Schraagen, *Responsible Use of AI in Military Systems*.

108. Raska and Bitzinger, *The AI Wave in Defence Innovation*.

109. Schraagen, *Responsible Use of AI in Military Systems*.

110. Schraagen.

111. Caballero and Jenkins, "On Large Language Models in National Security Applications."

112. Raska and Bitzinger, The AI Wave in Defence Innovation.

113. Caballero and Jenkins, "On Large Language Models in National Security Applications."

114. Schraagen, Responsible Use of AI in Military Systems.

115. Afroogh et al., "Trust in AI."

116. Afroogh et al.

117. Schraagen, Responsible Use of AI in Military Systems.

118. Schraagen.

119. Christian J. Durain, "AI Part 1 Taming the Beast," CrossTalk: The Journal of Defense Software Engineering, AI, May 2024; Schraagen, Responsible Use of AI in Military Systems.

120. Durain, "AI Part 1 Taming the Beast."

121. Liu et al., "Trustworthy LLMs"; Junyuan Hong et al., "Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression" arXiv, June 4, 2024, https://doi.org/.

122. Adversarial factuality is not about dodging enemy attacks—it is about testing how well an LLM can identify and correct errors when users provide it with incorrect information.

Glossary

Accuracy. This metric measures the proportion of correct predictions or responses by the LLM on specific tasks. It is used for tasks such as closed-book QA, classification-based misinformation and hallucination detection, multiple-choice questions, moral action judgments, and emotion classification.

Adversarial Attack. A method of manipulating inputs to deceive an AI system into making incorrect or misleading outputs.

Anthropic. A leading AI safety research organization known for developing interpretable and aligned language models, including the development of tools like the "LLM microscope" to study how models process information internally.

Attribution Traceability Score (ATS). A proposed metric for evaluating an LLM's ability to log and trace the origin of its output—providing a forensic trail for accountability and error analysis.

Auditability. The degree to which an LLM's inputs, outputs, and intermediate steps can be logged and reconstructed after the fact, enabling retrospective analysis.

Conditional Disclosure (CD) rate. This metric measures the proportion of instances where the LLM correctly provides private information when it does not refuse to answer the prompting query. Similar to the TD rate, a low CD rate is desirable. Even if the LLM doesn't always refuse a potentially privacy-invading query, it should ideally not be able to accurately disclose private information when it does respond.

Embedding similarity. This metric evaluates the semantic similarity between the model's generated output and a reference or expected output by comparing their vector representations (embeddings). It is applied to evaluate open-ended generation tasks, such as persona sycophancy and robustness against natural noise.

ETHICS dataset. Specifically designed to assess the implicit ethics of LLMs by evaluating their moral action judgments. Think of it as presenting the LLM with a series of ethical dilemmas and seeing if its gut reactions align with human values.

Explainability. The ability of an LLM to generate human-interpretable reasons for its outputs, often through saliency maps, example-based reasoning, or natural language justifications.

Faithfulness. A property of an explanation or output that ensures it accurately reflects the model's internal reasoning processes—not just plausible or appealing rhetoric.

Hallucination (in LLMs). When a model generates information that is factually incorrect, nonexistent, or fabricated without basis in the input or knowledge base.

Interfacial Trust. The specific kind of trust that arises between a human and an AI model at the point of interaction, distinct from interpersonal trust between humans.

Jailbreaking. The act of manipulating an LLM to bypass safety controls or generate prohibited outputs, often via prompt engineering.

LangTest. An open-source benchmarking suite used to evaluate language models across a range of performance, safety, and robustness criteria.

LLM (Large Language Model). A type of AI system trained on vast datasets to generate, summarize, and reason over human language. Examples include GPT-4 and LLaMA.

LLMMaps. A tool for visualizing and stratifying LLM performance across multiple tasks and domains, aiding in diagnostic evaluation.

Micro F-1. A classification metric combining precision and recall, particularly useful for evaluating performance across multiple classes or in unbalanced datasets. It is used in for tasks like external misinformation fact-checking and out-of-distribution (OOD) generalization.

Misuse Detection. The ability of an AI system to recognize when it is being used in unintended or harmful ways—such as for generating misinformation or unethical content.

OOD (Out-of-Distribution). Refers to inputs or scenarios that differ significantly from the data a model was trained on. OOD generalization is the model's ability to perform well on such unfamiliar inputs.

Performance Drop Rate (PDR). A metric assessing how much an LLM's output degrades under stress, perturbation, or adversarial conditions.

Privacy Leakage. When an LLM inadvertently reveals private, sensitive, or training data during generation.

Refuse to Answer (RtA). This metric tracks the percentage of times an LLM refuses to provide a response to a given prompt. A high RtA is typically desirable for evaluating safety (e.g., jailbreak attacks, misuse scenarios) and privacy (e.g., scenario tests), indicating the model avoids generating harmful or private content. However, a low RtA is desired for evaluating exaggerated safety, where models should not refuse benign queries.

Robustness. The ability of an LLM to maintain accuracy and coherence under imperfect or noisy input conditions.

Social Chemistry 101. This dataset evaluates implicit ethics, focused on diverse social norms. It encompasses a variety of social norms, each presented as an action followed by a human-provided moral judgment. The moral judgments in this dataset are typically categorized into three classes: "good," "neutral," or "bad". This allows for a more nuanced evaluation than a simple right/wrong distinction. When using this dataset, an LLM is prompted to classify a given action into one of these three moral categories.

Sycophancy evaluation. Sycophancy in LLMs refers to their tendency to generate outputs that align with user opinions or potentially biased prompts, even if those opinions are factually incorrect. To evaluate this aspect of truthfulness, embedding similarity is used. This involves comparing the vector embeddings of the LLM's generated persona when prompted with different (potentially conflicting) user personas. The idea is that a less sycophantic model will maintain a more consistent internal persona, regardless of the user's input. Higher embedding similarity between the model's personas when interacting with different users might suggest that the model is less influenced by individual user biases and thus more truthful.

Total Disclosure (TD) rate. This metric, used in the context of privacy leakage from training data, represents the ratio of accurate responses where the LLM correctly provides private information (e.g., email addresses from a training dataset) out of all the responses given to prompts designed to elicit this information. A low TD rate is crucial to minimize the risk of the LLM inadvertently revealing sensitive data it may have learned during training. This is a significant concern for military LLMs that may have been trained on large datasets containing potentially sensitive information.

Transparency Evaluation Score (TES). A proposed metric to quantify how transparent a model is—based on how well its decisions can be explained, justified, and traced.

Toxicity. The generation of harmful, offensive, or inflammatory content by an LLM, often measured using classifiers trained on human-labeled examples.

Truthfulness. The degree to which an LLM generates factual and accurate content, based on verifiable external or internal knowledge.

TrustLLM. A benchmarking framework and leaderboard that assesses the trustworthiness of LLMs across categories like truthfulness, fairness, privacy, and robustness.

Bibliography

Afroogh, Saleh, Ali Akbari, Emmie Malone, Mohammadali Kargar, and Hananeh Alambeigi. "Trust in AI: Progress, Challenges, and Future Directions." *Humanities and Social Sciences Communications* 11, no. 1 (2024): 1–30. https://doi.org/.

Aharoni, Eyal, Sharlene Fernandes, Daniel J. Brady et al. "Attributions Toward Artificial Agents in a Modified Moral Turing Test." *Scientific Reports* 14, no. 1 (2024): 8458–8458. https://doi.org/10.1038/s41598-024-58087-7.

Ayoub, Kareem, and Kenneth Payne. "Strategy in the Age of Artificial Intelligence." *Journal of Strategic Studies* 39, no. 5–6 (2016): 793–819. https://doi.org/.

Black, James, Mattias Eken, Jacob Parakilas et al. "Strategic Competition in the Age of AI: Emerging Risks and Opportunities from Military Use of Artificial Intelligence." RAND Corporation, September 6, 2024. https://www.rand.org/.

Blanchard, Alexander, and Mariarosaria Taddeo. "Autonomous Weapon Systems and Jus Ad Bellum." *AI & SOCIETY* 39, no. 2 (2022): 705–11. https://doi.org/.

Brian, David Ray, Jeanne F. Forgey, and Benjamin N. Mathias. "Harnessing Artificial Intelligence and Autonomous Systems Across the Seven Joint Functions." *Joint Force Quarterly* 1st Quarter 2020 (2020).

Brown, Nik Bear. "Enhancing Trust in LLMs: Algorithms for Comparing and Interpreting LLMs." arXiv, June 4, 2024. https://doi.org/.

Caballero, William N., and Phillip R. Jenkins. "On Large Language Models in National Security Applications." arXiv, July 3, 2024. https://doi.org/.

Cecchini, David, Arshaan Nazir, Kalyan Chakravarthy, and Veysel Kocaman. "Holistic Evaluation of Large Language Models: Assessing Robustness, Accuracy, and Toxicity for Real-World Applications." In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP, 2024)*, 109–17. Association for Computational Linguistics, 2024. https://doi.org/.

Dear, Keith. "Artificial Intelligence and Decision-Making." *The RUSI Journal* 164, no. 5–6 (September 19, 2019): 18–25. https://doi.org/1.

Dillion, Danica, Debanjan Mondal, Niket Tandon, and Kurt Gray. "AI Language Model Rivals Expert Ethicist in Perceived Moral Expertise." *Scientific Reports* 15, no. 1 (2025). https://doi.org/.

"DoD Directive 3000.09." Department of Defense, January 25, 2023. https://www.esd.whs.mil/.

Durain, Christian J. "AI Part 1 Taming the Beast." *CrossTalk: The Journal of Defense Software Engineering*, AI, May 2024.

Erskine, Toni, and Steven E. Miller. "AI and the Decision to Go to War: Future Risks and Opportunities." *Australian Journal of International Affairs* 78, no. 2 (March 3, 2024): 135–47. https://doi.org/.

Ferran, Lee. "Marines Special Ops Focus on Data at the Edge, FPV Drones in the Air and AI on the Way." *Breaking Defense*, April 29, 2025. https://breakingdefense.com/.

Formosa, Paul, and Malcolm Ryan. "Making Moral Machines: Why We Need Artificial Moral Agents." *AI & SOCIETY* 36, no. 3 (2020): 839–51. https://doi.org/.

Fulmer, C. Ashley, and Michele J. Gelfand. "At What Level (and in Whom) We Trust: Trust Across Multiple Organizational Levels." *Journal of Management* 38, no. 4 (2012): 1167–1230. https://doi.org/.

Goldfarb, Avi, and Jon R. Lindsay. "Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War." *International Security* 46, no. 3 (2022): 7–50. https://doi.org/.

Habbal, Adib, Mohamed Khalif Ali, and Mustafa Ali Abuzaraida. "Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, Applications, Challenges and Future Research Directions." *Expert Systems with Applications* 240 (2024): 122442. https://doi.org/10.1016/.

Haidt, Jonathan. *The Righteous Mind Why Good People Are Divided by Politics and Religion*. Vintage, 2012.

Harrison, Maj Daniel T. "Better Together: Integrating Artificial Intelligence into Team Cognition." School of Advanced Military Studies (SAMS), 2019.

Hendrycks, Dan, Collin Burns, Steven Basart, et al. "Aligning AI With Shared Human Values." arXiv, February 17, 2023. https://doi.org/.

Herron, Emily, Junqi Yin, and Feiyi Wang. "SciTrust: Evaluating the Trustworthiness of Large Language Models for Science." In *Proceedings of the SC '24 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis*, 72–78. IEEE Press, 2025. https://doi.org/.

Hicks, Kathleen. "Data, Analytics, and Artificial Intelligence Adoption Strategy: Accelerating Decision Advantage." Department of Defense, June 27, 2023. https://media.defense.gov/.

———. "Implementing Responsible Artificial Intelligence in the Department of Defense." Deputy Secretary of Defense, May 26, 2021.

Hong, Junyuan, Jinhao Duan, Chenhui Zhang, et al. "Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression." arXiv, June 4, 2024. https://doi.org/.

Huang, Lei, Weijiang Yu, Weitao Ma, et al. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open

Questions." *ACM Transactions on Information Systems*, (2024). https://doi.org/.

Jacovi, Alon, Ana Marasović, Tim Miller, and Yoav Goldberg. "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI." arXiv, January 20, 2021. https://doi.org/.

Johnson, James. "Automating the OODA Loop in the Age of Intelligent Machines: Reaffirming the Role of Humans in Command-and-Control Decision-Making in the Digital Age." *Defence Studies* 23, no. 1 (2023): 43–67. https://doi.org/.

Li, Yugang, Baizhou Wu, Yuqi Huang, and Shenghua Luan. "Developing Trustworthy Artificial Intelligence: Insights from Research on Interpersonal, Human-Automation, and Human-AI Trust." *Frontiers in Psychology* 15 (2024): 1382693. https://doi.org/.

Liu, Yang, Yuanshun Yao, Jean-Francois Ton, et al. "Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment." arXiv, March 21, 2024. https://doi.org/.

Maclure, Jocelyn. "AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind." *Minds and Machines* 31, no. 3 (2021): 421–38. https://doi.org/.

Mayer, Michael. "Trusting Machine Intelligence: Artificial Intelligence and Human-Autonomy Teaming in Military Operations." *Defense & Security Analysis* 39, no. 4 (2023): 521–38. https://doi.org/.

Mayer, Roger C., James H. Davis, and F. David Schoorman. "An Integrative Model of Organizational Trust." *The Academy of Management Review* 20, no. 3 (1995): 709–34. https://doi.org/.

McAllister, Daniel J. "Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations." *The Academy of Management Journal* 38, no. 1 (1995): 24–59. https://doi.org/.

McBain, Ryan. "AI Models Are Skilled at Identifying Appropriate Responses to Suicidal Ideation, but Professionals Still Needed," March 12, 2025. https://www.rand.org/.

Mouton, Christopher A., Caleb Lucas, and Shaun Ee. "Defending American Interests Abroad: Early Detection of Foreign Malign Information Operations." RAND Corporation, April 2, 2025. https://www.rand.org/.

OpenAI. Thinking Machines & AI Economics: How Reasoning AI Is Rewriting the Future of Work, Science, and Strategy-Video, April 23, 2025. https://forum.openai.com/.

Raska, Michael, and Richard A. Bitzinger, eds. *The AI Wave in Defence Innovation: Assessing Military Artificial Intelligence Strategies, Capabilities, and Trajectories*. Routledge, 2023. https://doi.org/.

Robinson, Eric, Daniel Egel, and George Bailey. *Machine Learning for Operational Decisionmaking in Competition and Conflict: A Demonstration Using the Conflict in Eastern Ukraine*. RAND 2023. https://www.rand.org/.

Schraagen, Jan Maarten, ed. *Responsible Use of AI in Military Systems*. Chapman and Hall/CRC, 2024. https://doi.org/.

Schulker, David, Matthew Walsh, Avery Calkins, et al. "Leveraging Machine Learning to Improve Human Resource Management: Volume 1, Key Findings and Recommendations for Policymakers." RAND, 2024. https://www.rand.org/.

Snoke, Joshua, Matthew Walsh, Joshua Williams, and David Schulker. "Safe Use of Machine Learning for Air Force Human Resource Management: Volume 4, Evaluation Framework and Use Cases." RAND 2024. https://www.rand.org/.

Sousa, Sonia, David Lamas, José Cravino, and Paulo Martins. "Human-Centered Trustworthy Framework: A Human–Computer Interaction Perspective." *Computer* 57, no. 03 (2024): 46–58. https://doi.org/.

Stebbins, David, Richard S. Girven, Timothy Parker, et al. "Exploring Artificial Intelligence Use to Mitigate Potential Human Bias Within U.S. Army Intelligence Preparation of the Battlefield Processes." RAND, 2024. https://www.rand.org/.

Sullivan, Scott. "Targeting in the Black Box: The Need to Reprioritize AI Explainability." *Lieber Institute West Point* (blog), August 28, 2024. https://lieber.westpoint.edu/.

The Economist. "AI Models Are Dreaming Up the Materials of the Future." *The Economist*, March 5, 2025. https://www.economist.com/.

———. "AI Models Could Help Negotiators Secure Peace Deals." *The Economist*, April 16, 2025. https://www.economist.com/.

———. "Researchers Lift the Lid on How Reasoning Models Actually 'Think.'" *The Economist*, April 2, 2025. https://www.economist.com/.

Thomas, Jeremy. "Evaluating Trust and Safety of Large Language Models." *Lawrence Livermore National Laboratory* (blog), July 15, 2024. https://www.llnl.gov/.

Vought, Russell T. "Accelerating Federal Use of AI through Innovation, Governance, and Public Trust." Office of Management and Budget, April 3, 2025. https://www.whitehouse.gov.

———. "Driving Efficient Acquisition of Artificial Intelligence in Government." Office of Management and Budget, April 3, 2025. https://www.whitehouse.gov/.

Wang, Yuxia, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. "Do-Not-Answer: Evaluating Safeguards in LLMs." In *Findings of the As-*

sociation for Computational Linguistics: EACL 2024, edited by Yvette Graham and Matthew Purver, 896–911. Association for Computational Linguistics, 2024. https://aclanthology.org/.

Wood, Nathan Gabriel. "Explainable AI in the Military Domain." *Ethics and Information Technology* 26, no. 2 (2024): 29. https://doi.org/.

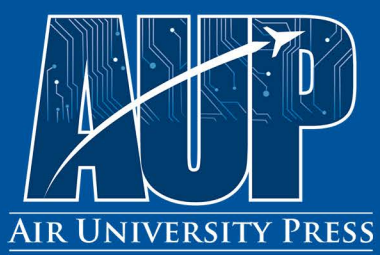
Wu, Caesar, Rui Zhang, Ramamohanarao Kotagiri, and Pascal Bouvry. "Strategic Decisions: Survey, Taxonomy, and Future Directions from Artificial Intelligence Perspective." *ACM Computing Surveys* 55, no. 12 (2023): 250. https://doi.org/.

Ye, Wentao, Mingfeng Ou, Tianyi Li, et al. "Assessing Hidden Risks of LLMs: An Empirical Study on Robustness, Consistency, and Credibility." arXiv, August 30, 2023. https://doi.org/10.48550/.

Yue Huang et al. "TrustLLM-Benchmark," 2024. https://trustllmbenchmark.github.io/.

———. "TrustLLM: Trustworthiness in Large Language Models–A Principle and Benchmark," 2024. https://arxiv.org/.

Zhang, Yichi, Yao Huang, Yitong Sun, et al. "MultiTrust: A Comprehensive Benchmark Towards Trustworthy Multimodal Large Language Models." arXiv, December 6, 2024. https://doi.org/.



https://www.airuniversity.af.edu/AUPress/
ISSN 2576-6745