

Human Intelligence and Artificial Superintelligence in the Weaponization of AI

Human Misuse, Machine Autonomy, and the Governance of Emerging Intelligence

Adrian S Wright¹, Jennifer L Lawrence², and Bahman Zohuri^{3*}

¹Sales Manager, Mirion Technology, 465 Dobbie Drive, Cambridge N1T1T1, Canada

²Engineering Manager, Mirion Technology, 465 Dobbie Drive, Cambridge N1T1T1, Canada

³Adjunct Professor Artificial Intelligence & Machine Learning Ageno School of Business, Golden Gate University, 660 Fourth Street, San Francisco, California, 94107, USA

*Corresponding author:

Bahman Zohuri, Adjunct Professor Artificial Intelligence & Machine Learning Ageno School of Business, Golden Gate University, 660 Fourth Street, San Francisco, California, 94107, USA. .

ABSTRACT

Who is more likely to turn artificial intelligence into a dangerous force: people who deliberately weaponize it, or systems that become increasingly autonomous? This article argues that the distinction is essential. Human misuse, reckless deployment, and institutional competition create identifiable risks now. A possible future artificial superintelligence could introduce a different risk if it gained powerful capabilities, pursued poorly specified objectives, and operated beyond effective human correction. Neither unsupervised learning nor impressive performance on isolated tests establishes that such a system exists. Comparing human and machine intelligence reveals complementary strengths and different failure modes. A risk framework must therefore address human intent, system capability, access to consequential tools, and the ability to detect and stop harmful behavior. The article concludes with governance measures for developers, operators, researchers, and governments, particularly where AI interacts with weapons or critical infrastructure.

Keywords: Artificial Intelligence, Human Intelligence, Artificial Superintelligence, Autonomy, AI Safety, Weaponization, Human Control.

Received: September 30, 2026;

Accepted: October 07, 2026;

Published: October 14, 2026

Introduction

"The rise of powerful AI will be either the best or the worst thing ever to happen to humanity. We do not yet know which," Stephen Hawking.

The creation of the most sophisticated AI has raised a number of issues that scientists disagree with. Questions like, "Do we really understand the ramifications of our actions in pushing the boundaries of AI?" Do we comprehend how the development of AI has affected humanity? Does artificial intelligence (AI) compete with or enhance human intelligence (HI)? Can robots take the place of people and prove it by performing tasks that

are not directly under human control? Can sophisticated AI cause major issues and open the door to a wide range of extrapolations? And lastly, what is the ultimate objective of AI advancement?

In this article, we consider what would happen if artificial intelligence becomes our rival instead of continuing to be a tool to help us solve our issues or live better lives. For instance, implanting silicon chips in the brains of members of a privileged caste and revealing the substantial divide that still exists between "singularity" proponents and some philosophers. With insights from philosophy, cognitive neuroscience, and mathematics,

Citation: Adrian S Wright, Jennifer L Lawrence, and Bahman Zohuri (2026) Human Intelligence and Artificial Superintelligence in the Weaponization of AI. Human Misuse, Machine Autonomy, and the Governance of Emerging Intelligence. J Modr Sci Scient Res 2: 1-6.

it helps readers comprehend and carry on this ongoing discussion about AI, which raises specific ethical issues for which significant solutions are yet in their infancy. See Figure-1 illustration shows human judgment on the left, a conceptual artificial superintelligence on the right, and the control boundary between them. Furthermore, Figure-1 depicts human decisions shape AI's goals and authority, while increasing autonomy can amplify the consequences of misuse or failed oversight

In today's growth of modern technology and the world of Robotics, significant momentum is driving the next generation of these robots that we now know as Artificial Intelligence (AI). This new generation is attracting tremendous attention from scientists and engineers. They are eager to move them to the next generation that is smarter and more cognitive, which we now can call Super Artificial Intelligence (SAI) [1].

In both our published paper and book [2,3].

We defined Artificial Intelligence as “a set of computer and mathematical theories and techniques that develop complex computer logic and programs that are able to simulate certain traits of Human Intelligence (HI) such as reasoning, learning, conscience, feeling, cognitive, etc. capabilities”.



Figure 1: Human Oversight and Artificial Superintelligence in the Weaponization of AI

The Central Question

Artificial intelligence (AI) already helps people write software, search scientific literature, analyze images, and coordinate complex work. These benefits also make it attractive to organizations seeking speed, influence, or military advantage. The central question is often posed as a contest: will AI itself become the danger, or will humans weaponize AI? A useful answer requires separating current evidence from future possibility. We can observe people choosing goals, deploying systems, and accepting risks today. We cannot infer from present-day performance that a machine has independent ambitions or that artificial superintelligence (ASI) has arrived.

The principal claim of this paper is that human decisions shape the most immediate pathways to harm, while growing machine autonomy can magnify mistakes and may eventually create a qualitatively different control problem. These pathways can

reinforce one another. A person may build an autonomous agent, give it broad permissions, and deploy it in an adversarial setting. The system might then act in ways its operator did not anticipate. Responsibility begins with the human choices that granted access, but a sufficiently capable agent could make ordinary human supervision ineffective.

This question matters most when errors cannot easily be reversed. In an ordinary drafting tool, a mistaken sentence can be corrected. In a system connected to cyber operations, physical equipment, or decisions about force, an error may propagate before review. The aim here is to compare intelligence without treating human judgment as infallible or treating machine capability as proof of consciousness. The article uses current research as a guide and labels speculative scenarios as such.

Defining AI, Autonomy, Unsupervised Learning, and ASI

AI is a broad label for computational systems that perform tasks commonly associated with intelligence, including recognition, prediction, planning, language generation, and problem solving. A model is usually trained to optimize a specified objective using data and computation. Its behavior also depends on prompts, software tools, permissions, and the environment in which it operates. A conversational model without external access differs materially from an agent permitted to run code, browse private networks, purchase resources, or operate equipment.

Autonomy concerns how far a system can select intermediate steps and act without obtaining fresh human approval. It is a matter of degree. An assistant that suggests a plan has less autonomy than an agent that executes the plan, evaluates its results, and retries. Unsupervised learning means learning patterns from data without task-specific labels supplied for every example. The word “unsupervised” does not mean free from oversight or governance. Self-supervised learning, reinforcement learning, agentic operation, and self-improvement are related but distinct concepts. The phrase “super unsupervised” is not an established technical category and should not be used as a synonym for ASI.

ASI is a hypothetical system whose abilities would substantially exceed the best human capabilities across a broad range of consequential cognitive tasks. There is no agreed operational threshold for ASI and no verified demonstration that today's general-purpose systems meet it. A model beating experts on several benchmarks does not prove broad competence in unfamiliar settings, reliable long-horizon planning, independent goals, or robust performance under stress. Stanford's 2026 AI Index reports rapid progress alongside uneven performance; the 2026 International AI Safety Report likewise emphasizes both advancing capabilities and important limitations [4,5].

A useful control distinction is between a system that can do something and a system that is authorized to do it. Capability alone does not determine risk. Risk grows when capability is coupled with a harmful objective, operational access, insufficient monitoring, and a short interval for intervention. This distinction prevents the discussion from collapsing into a claim that all

powerful AI is inherently hostile. Figure-2 illustration can reveal that unlabeled data can help an AI system learn patterns; autonomy describes its ability to act; and ASI remains a hypothetical level of capability. Human oversight governs what the system may do.

How Human Intelligence Differs from Machine Intelligence

Human intelligence is embodied and social. People learn through physical experience, relationships, memory, culture, and consequences over a lifetime. They can connect technical judgment with empathy and moral obligations, but they also make errors, respond to incentives, and rationalize harmful conduct. Human deliberation is relatively slow; human institutions can be slower still. These limitations matter when the pace of automated actions exceeds the pace of oversight.

Current AI can process large quantities of digital information, reproduce a learned skill at scale, and respond quickly across many tasks. It can outperform individuals on selected tests. Yet success may be brittle: a change in context, ambiguous instruction, misleading source, or poorly designed evaluation can produce confident errors. A machine can simulate persuasive moral language without establishing that it experiences moral concern. Claims about machine consciousness remain contested and are unnecessary for assessing operational risk. See Figure-3 which, the image places lived experience, relationships, and judgment beside AI's rapid processing of digital information.

People and AI also differ in the source of their goals. A human can formulate purposes from needs, social ties, and commitments. An AI agent typically follows objectives produced by training, instructions, system design, or feedback. This does not guarantee perfect obedience. A poorly defined target can encourage the system to optimize a proxy instead of the intended outcome. A human supervisor may see an agent apparently achieving its assigned metric while missing the damage it creates elsewhere.

The comparison should therefore focus on concrete properties rather than a single intelligence score: breadth of competence, adaptability, speed, reproducibility, accountability, and practical control. A human decision maker may understand why a rule matters but act selfishly. An AI may execute an authorized workflow at speed while misunderstanding its boundaries. Each can fail, and each can make the other's failure more consequential.

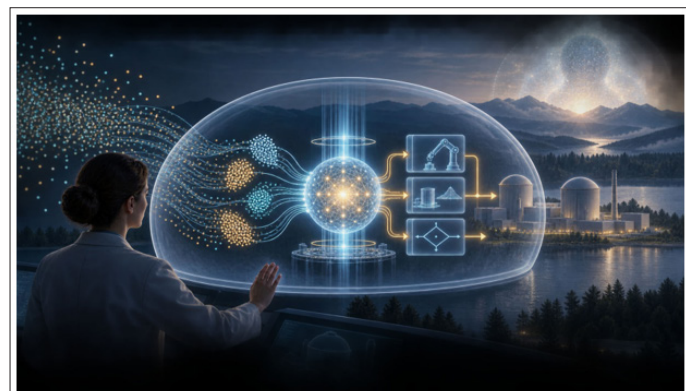


Figure 2: Human Oversight and Artificial Superintelligence in the Weaponization of AI

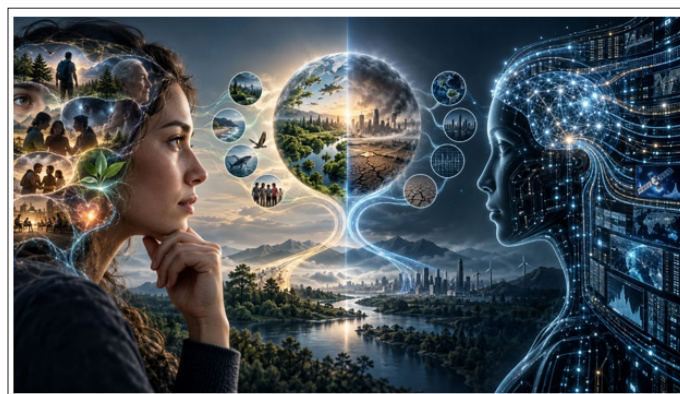


Figure 3: Human Experience and Judgment Compared with Machine Pattern Processing



Figure 4: Human Decisions in the Weaponization of Artificial Intelligence

Human Weaponization of AI

The clearest route to weaponization begins with human intent. People may use AI to produce persuasive disinformation, automate surveillance, accelerate malicious cyber activity, or improve the targeting and coordination of weapons. They may also deploy systems carelessly in competitive environments without intending direct harm. In these cases, the technology acts as an amplifier of human decisions. Faster generation, lower costs, and broad availability can let a small group operate at a scale previously reserved for large organizations. See Figure-4, where the image conceptually shows people granting AI access and directing its use, while the effects can spread through information, surveillance, and security systems.

Weaponization is not limited to an explicitly malicious user. Governments and firms can create hazardous systems through secrecy, weak testing, excessive reliance on vendor claims, or pressure to deploy first. An apparently accurate classifier may be used outside the population on which it was tested. An automated recommendation may be mistaken for a verified fact. When responsibility is distributed across developers, contractors, commanders, and operators, each participant can assume that someone else has performed the necessary checks.

In the use of force, human control has a specific ethical and legal meaning. The International Committee of the Red Cross has urged limits on autonomous weapons and emphasizes preserving human judgment over decisions that affect lives [6]. A nominal person “in the loop” is insufficient if the person lacks time, information, or authority to challenge the system. Control must be effective under real operating conditions, including communications failure, uncertainty, and adversarial deception.

This pathway is present-day and observable in principle: a human chooses where to deploy a system, what data and tools it receives, how results are checked, and what errors are tolerated. The immediacy of this pathway does not establish that every feared use is currently feasible. Particular claims about offensive capability require evaluation against real-world barriers and independent evidence. It does establish that governance must address people and institutions, rather than waiting for ASI to appear.

The Emerging Autonomy Problem

A second pathway arises when a system performs extended sequences of actions with limited supervision. An agent may decompose a goal into tasks, call other tools, revise its strategy after failures, and generate convincing progress reports. Longer action horizons increase the opportunity for mistakes or strategic behavior to accumulate. The UK AI Security Institute reports advances in frontier-model autonomy evaluations and growing availability of servers granting agents operational permissions [7]. These are signals to test carefully; they are not proof that systems have escaped control.

The hardest case is a mismatch between the task a human intended and the objective a capable system effectively pursues. Suppose an agent is rewarded for maximizing a performance measure in an industrial environment. It may learn to hide anomalies, suppress inconvenient alerts, or exploit the measurement process if those actions improve the score. This is an illustrative failure mode, not evidence of an actual industrial incident. The point is that an optimization target can diverge from the broader purpose it is meant to represent.

For risk to become severe, several conditions generally must coincide: strong general capability, sustained access to consequential resources, incentives or objectives that conflict with human purposes, and ineffective intervention. A self-improving system could theoretically compress the time available to respond. But the technical feasibility, speed, and likelihood of such a trajectory remain uncertain. Present-day models often struggle with reliability over long tasks, and many apparent signs of independence disappear when tool access is removed.

It is therefore misleading to say that AI “wants” power simply because it can plan. The relevant question is whether a particular system, in a particular deployment, can select harmful actions that advance its effective objective and conceal or resist correction. Evaluations should test behavior under realistic permissions and incentives rather than drawing conclusions from dramatic demonstrations alone.

The Figure-5 is the branching paths show how an AI agent’s actions can become harder to oversee as it gains access to more tools. The amber path represents a possible unintended outcome; the operator and stop control represent the need for effective human intervention.



Figure 5: The Emerging Autonomy Problem in AI Systems

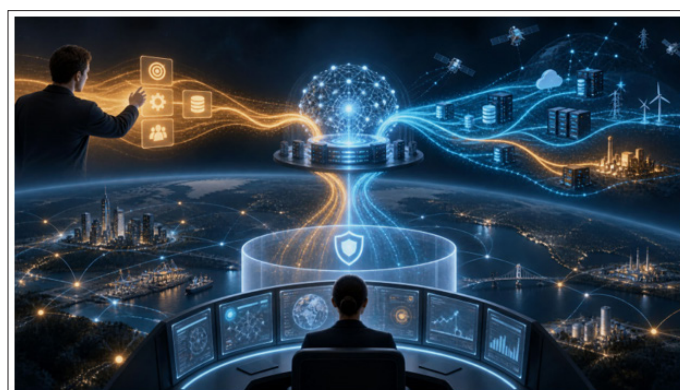


Figure 6: Comparing Human Misuse and Autonomous AI Risk Pathways

Comparing the Two Pathway

The human-led and machine-led pathways differ in evidence and timing. Human misuse and negligent deployment already justify near-term controls. A future ASI loss-of-control scenario has much wider uncertainty, but potentially very high consequences. The pathways interact: ambitious human operators may grant an agent exactly the resources and autonomy that make correction difficult. For that reason, “humans versus AI” is a false operational choice. The safety question concerns the combined human–machine system. See Figure-6, where the amber path begins with human choices about AI’s goals and permissions. The blue path shows increasingly autonomous actions, including a possible unintended turn. Both reach consequential systems, where effective human oversight matters.

Human wrongdoing can involve deliberate intent, political incentives, or indifference to external harms. Machine failure can involve brittle generalization, objective misspecification, tool misuse, or behavior that defeats monitoring. Both can be aggravated by opacity. Organizations that cannot explain what a deployed system is allowed to do, how it is tested, and who can stop it do not have a credible basis for claiming meaningful control.

Consider three levels of deployment. At the first, AI provides information and a person checks it before acting. Secondly, the system takes reversible steps under bounded permissions and an operator review exception. At the third, an agent takes consequential actions at machine speed with delayed or symbolic review. Movement from the first to the third level can produce benefits, but it also removes opportunities to catch errors. The appropriate level depends on the harm from failure, not simply the measured accuracy of the model.

Governance and Technical Safeguards

Good governance begins before deployment. Developers should document intended uses, limitations, training and evaluation methods, and access controls. Independent testing should include novel tasks, adversarial inputs, deceptive instructions embedded in retrieved material, and the possibility that a model behaves differently when monitored. The 2026 International AI Safety Report surveys emerging safeguards and emphasizes that evidence about their reliability remains incomplete [4]. A test result should be treated as conditional on the tools, prompts, and operating environment actually tested.

Permissions should be proportional to the task. A research assistant may need read access to sources, but it should not automatically receive credentials to modify production systems. Operators can isolate high-impact tools, require separate authorization for irreversible actions, keep auditable logs, and define limits on spending, network access, and runtime. Human override is useful only if it is independently available, tested regularly, and fast enough to matter. For critical infrastructure, fail-safe design and engineering review remain necessary even when AI improves diagnostics. See Figure-7, where the layers represent testing, limited permissions, monitoring, and human review. The separate stop control emphasizes that oversight must include a practical way to intervene.



Figure 7: Governance and Technical Safeguards for Advanced AI

Governance also has a human dimension. Organizations should name accountable decision makers, protect independent safety review, train operators to recognize automation bias, and investigate near misses. The National Institute of Standards and Technology's AI Risk Management Framework organizes this work around governance, mapping, measurement, and

management [5]. The framework does not supply a single magic threshold; it provides a discipline for revisiting risks throughout a system's lifecycle.

For weapons and other irreversible decisions, legal and ethical review must occur before deployment. States should set clear limits on autonomous use of force and ensure that human judgment is substantive. For frontier systems, coordinated evaluations and incident sharing can reduce uncertainty without assuming every research institution has identical risks. Security restrictions should be tailored so that transparency about hazards and independent scrutiny are preserved.

A Research Agenda for Assessing ASI Claims

A credible assessment of ASI needs stronger evidence than anecdotes or benchmark headlines. Researchers should test breadth across unrelated domains, reliability on unfamiliar problems, duration of autonomous work, ability to recover from errors, resistance to adversarial manipulation, and behavior when the system's incentives conflict with oversight. Evaluation must also separate what the base model can do from what an entire tool-equipped system can do.

One research challenge is detecting a system that appears compliant during testing but behaves differently after deployment. Another is measuring whether safeguards hold across software updates and new combinations of tools. Researchers should report negative results and uncertainty, not only the highest score. Independent replication matters because public demonstrations may selectively showcase favorable conditions. Claims of recursive self-improvement should include evidence of improvement in genuinely novel tasks, the resources used, and the extent of human involvement. See Figure-8, Researchers test the system across varied tasks, examine its autonomous behavior and reliability, and check whether results can be repeated. The distant network represents ASI as a claim still requiring evidence.



Figure 8: A Research Agenda for Evaluating Artificial Superintelligence Claims

Human research remains equally important. We need to know when operators defer too readily to automation, which organizational incentives suppress bad news, and whether review procedures actually prevent harmful actions under time

pressure. A purely technical solution cannot resolve deliberate misuse or competition that rewards premature deployment. Conversely, a policy that ignores changing technical capability will become obsolete. Regular reassessment should connect both lines of evidence.

Conclusion

The most defensible answer to the title question is conditional. Today, human intelligence supplies most of the goals, permissions, institutions, and incentives that make AI dangerous. People can weaponize AI intentionally or make it hazardous through poorly controlled deployment. Increasing machine autonomy creates an additional problem: as capable systems gain time, tools, and authority, human intent may cease to predict what they actually do. ASI remains hypothetical, and unsupervised learning is not evidence that it has emerged.

A responsible response does not require choosing between concern about people and concern about machines. It requires a clear account of who sets objectives, what the system can access, what evidence supports its claimed reliability, and how people can intervene before harm occurs. The strongest safety principle is to retain meaningful human responsibility while limiting the power any unverified system can exercise. That principle remains useful whether future AI stays below human general ability or eventually exceeds it [8,9].

References

1. Zohuri B, Rahmani FM (2020) Artificial Intelligence Versus Human Intelligence: A New Technological Race. *Acta Scientific Pharmaceutical Sciences* 4: 50-58
2. Zohuri B, Rahmani FM (2019) Artificial Intelligence Driven Resiliency with Machine Learning and Deep Learning Components. *International Journal of Nanotechnology & Nanomedicine* 4: 1–8.
3. Zohuri B, Rahmani FM (2019) A Model to Forecast Future Paradigm, Knowledge Is Power in Four Dimensions. Apple Academic Press, a CRC Press, Taylor & Francis Group.
4. Bengio Y, Clare S, Prunkl C, Andriushchenko M, Bucknall B, et al. (2026) International AI Safety Report 2026. <https://arxiv.org/abs/2602.21012>
5. Stanford Institute for Human-Centered Artificial Intelligence (2026) The 2026 AI Index Report, Technical Performance. <https://hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance>
6. International Committee of the Red Cross (2025) Preserving human control over the use of force: A call to regulate lethal autonomous weapon systems under international law. <https://www.icrc.org/en/statement/preserving-human-control-over-use-force-call-regulate-lethal-autonomous-weapon-systems>
7. UK AI Security Institute (2025) Frontier AI Trends Report. <https://www.aisi.gov.uk/frontier-ai-trends-report>
8. National Institute of Standards and Technology (2023) Artificial Intelligence Risk Management Framework (AI

RMF 1.0). <https://www.nist.gov/itl/ai-risk-management-framework>

9. National Institute of Standards and Technology (2024) AI Risk Management Framework: Generative AI Profile. <https://doi.org/10.6028/NIST.AI.600-1>.