FUTURE THREATS AND RISKS - I: DEEPFAKES, AUTONOMOUS WEAPONS, AND AI SAFETY

Dr. Yasir Atalan
Center for Strategic and International Studies

Attribution: Atalan, Y. (2026). Future threats and risks - I: Deepfakes, autonomous weapons, and AI safety. In H. Arslan, A. Taşdelen, E. Kuzucu Hıdır, & O. A. Çetin (Eds.), *Artificial intelligence and national technology reflections* (pp. 559–582). Turkish Academy of Sciences. https://doi.org/10.53478/TUBA.978-625-6110-86-1.ch23

10.53478/TUBA.978-625-6110-86-1.ch23

FUTURE THREATS AND RISKS - I: DEEPFAKES, AUTONOMOUS WEAPONS, AND AI SAFETY

Dr. Yasir Atalan
Center for Strategic and International Studies

Abstract

Artificial intelligence (AI) is increasingly viewed as a general-purpose technology with the potential to reshape production, communication, and governance. At the same time, it is changing the threat landscape facing individuals, societies, and institutions. This chapter outlines future AI threats and risks using a two-by-two framework that distinguishes epistemic versus kinetic harms and misuse versus systemic sources. Epistemic harms destabilize the information environment and the conditions for forming reliable beliefs, while kinetic harms involve physical or cyber-physical damage. Misuse captures intentional, malicious deployment of AI; systemic risks arise from embedding AI in platforms, markets, and institutions even absent malign intent. The chapter then explores two case studies. First, deepfakes illustrate epistemic misuse by lowering the cost of credible deception and enabling the contestation of authentic evidence, thereby weakening public trust. Second, autonomous weapons and AI-enabled targeting pipelines illustrate kinetic risk by distributing lethal decision functions across human–machine ensembles, raising accountability challenges under international humanitarian law and increasing escalation pressures through speed and opacity. The chapter then reviews technical countermeasures, provenance standards, and emerging regulatory approaches, and explains why current responses lag behind rapid capability growth. It concludes by assessing how agentic AI may amplify and recombine these risks across domains.

Keywords

Artificial Intelligence, Deepfake, Autonomous Weapons, AI Governance, AI Safety

Introduction

Artificial intelligence is increasingly described as a general-purpose technology-potentially as transformative as electrification or the internal combustion engine. While "AI" often refers to specific machine-learning methods, recent advances in large language models (LLMs) have made it easier to combine multiple techniques into systems that can be directed through natural language. As these systems become embedded across sectors, the result is not just better tools, but new ways of organizing production, communication, and governance (Brynjolfsson & McAfee, 2017; DeSantis, 2024). In practical terms, the cognitive burden on users is now lower than ever. Many tasks now require less specialized expertise to plan, coordinate, and execute, as AI systems can interpret instructions and perform complex sequences of actions with minimal supervision.

Some accounts therefore place AI alongside earlier technological revolutions, emphasizing the possibility of deep economic and organizational reconfiguration (Karnofsky, 2016). The economic projections reflect this view. The global AI market is expected to grow from approximately $189 billion in 2023 to around $4.8 trillion by 2033 (UNCTAD, 2025). At the micro level, generative AI can increase task-level productivity in occupations such as customer support, coding, and consulting by 5 to more than 25 percent for many workers (Filippucci et al., 2024). Taken together, AI systems sit at the center of a new techno-economic paradigm that both promises productivity and reconfigures the distribution of power, wealth, and risk.

This transformative character cuts both ways. The same capabilities that support efficiency and innovation also generate novel threat vectors that can undermine human rights, political processes, social trust, and international security. More importantly, AI-related harms are not reducible to a single mechanism. They emerge through both the deliberate weaponization of AI by adversarial actors, and the unintended consequences of embedding AI into socio-technical systems at scale. To provide an analytical framework, this chapter develops a two-by-two matrix that maps AI-related threats along the axes of domain and source. The domain axis includes, first, epistemic threats which degrade and destabilize the information environment and the conditions under which individuals and societies form reliable beliefs. The other vector is kinetic threats which involve physical harm to bodies, infrastructures, or cyber-physical systems. The source axis, on the other hand, includes, first, threats arising from misuse, meaning intentional, adversarial deployment of AI to cause harm. The Second vector in this axis is systemic risks that arise from structural dependence on AI within platforms, markets, and institutions even in the absence of malicious intent. This framework yields four analytically distinct cells: epistemic–misuse (e.g., deepfakes, automated influence operations, AI-assisted social engineering), epistemic–systemic (e.g., engagement-maximizing

recommender systems, epistemic erosion and "liar's dividend" dynamics), kinetic–misuse (e.g., AI-enabled cyber-physical attacks, weaponized drones), and kinetic–systemic (e.g., accelerated targeting pipelines, algorithmically compressed decision cycles, accident-prone tightly coupled infrastructures).

The remainder of the chapter builds on this framework with two case studies. First, it analyzes deepfakes as an epistemic–misuse threat, tracing how synthetic content amplifies disinformation, destabilizes public trust and poses risks to the political institutions. Second, it examines autonomous weapons and AI-enabled targeting pipelines as a kinetic risk vector, emphasizing how autonomy operates on a continuum, how delegation of lethal functions raises accountability and compliance challenges under international humanitarian law, and how speed and opacity may increase escalation risks. For both cases, the chapter lays out the existing measures and governance responses and explains why they remain insufficient relative to the pace of technological change. It concludes by looking ahead to a threat landscape further complicated by emerging agentic AI capabilities-systems that can sense, reason, and act on behalf of users-potentially expanding both the scale and the complexity of future risks.

1. Mapping the Threat Landscape

Technology has always reshaped the threat landscape by changing what actors can do, what it costs to do it, and how visible or accountable the harm becomes. The printing press and mass media altered propaganda and censorship (Eisenstein, 1979; Briggs & Burke, 2009). Industrialization introduced lethal weapons, multiplied the destructive capacity, culminating in nuclear weapons. The digital revolution added cyberspace as a domain of conflict, allowing state and non-state actors to conduct espionage, sabotage, and psychological operations across borders at low cost. Artificial Intelligence builds on this history, but it does so by introducing distinctive properties. In the civilian space, advances in machine learning systems underpin recommender systems, predictive analytics, and generative tools that can write text, generate images, or control robots. Algorithmic decision systems in finance, policing, and welfare raise concerns about discrimination and opacity (Zarsky, 2016). Disinformation campaigns use social media and bots to hijack attention and polarize public discourse (Tucker et al., 2018). In military contexts, similar techniques support target recognition, anomaly detection, autonomous navigation, and automated decision support (Lindsay, 2020). What is new is not the existence of these threats, but the speed, personalization and realism with which they can be pursued.

The pathways that AI shape this threat landscape emerges out of the combination of two axes, namely domain and source. The domain dimension asks what the mode of threat is. On the one hand, epistemic threats damage the information environment

and the conditions under which individuals and societies can form reliable beliefs. On the other hand, kinetic risks posed by AI are considered to be threat vectors that could lead to fatal risks. AI systems provide critical risks in the kinetic/embodied domain in which they can affect physical space with AI capabilities. The other axis, namely the source axis, asks how these harms arise. While misuse covers malicious, intentional weaponization of AI by human actors, systemic risks refer to the risks that arise from the way AI is built into socio-technical structures, markets and institutions, even when no individual actor is explicitly trying to cause harm. This keeps the focus on AI as part of human organizational arrangements rather than as an external agent.

On the source side, the first threat pathway is intentional adversarial use. This is why discussions of AI threats are often framed in terms of "malicious use," "misuse," or "abuse" (Brundage et al., 2018; Blauth et al., 2022). The underlying assumption is that AI is not inherently a threat; rather, harm arises when humans deliberately employ AI to enhance hostile activities. In this view, AI primarily reduces the cost and increases the scale of familiar attack patterns. Actors can use AI to generate and spread disinformation, expand surveillance capacities in support of coercive governance, and enable social engineering through personalized deepfakes, impersonation, or voice cloning. In cyber operations, AI can automate vulnerability discovery to attack key nodes in critical infrastructure. AI tools similarly provide new capabilities for cyber attackers more generally. As societies grow more dependent on digital networks, these advantages become more strategically meaningful.

The same logic extends to biological and kinetic threats. Generative models and powerful search tools can, in principle, make it easier for individuals with harmful intentions to access information about pathogens and dissemination methods (Urbina et al., 2023). In military contexts, integrating AI into target identification and decision support creates new opportunities for adversaries to attack the data flows that the systems depend on. Training data can be poisoned, sensors can be spoofed and the broader information ecosystem can be flooded with synthetic or coordinated content in the aim of shaping perception and behavior. In all of these cases, AI serves as a force multiplier by providing new means for malicious actors to innovate their adversarial campaigns.

A second pathway involves systemic risks that arise even without adversarial intent. Focusing only on misuse obscures how widespread reliance on AI can shift authority, reorganize labor, and weaken accountability as AI becomes embedded in socio-technical systems (Roff et al., 2016). Here the concern is less about what hostile actors do, and more about how institutions behave when they depend on AI as an ordinary part of decision-making and operations (van de Poel, 2023). One obvious example is labor displacement through automation (DeCanio, 2016). Technological change has long produced productivity gains through creative destruction (Schumpeter, 1934).

Yet the prospect of highly capable, general-purpose AI renews the question of whether this wave of automation could generate qualitatively different forms of inequality and deskilling, and in extreme scenarios make large segments of human labor economically redundant. In these accounts, the risk is not only unemployment but a broader loss of agency in economic life (Ernst et al., 2019).

Loss of control also appears in safety-critical domains. When targeting, engagement, or safety decisions are delegated to AI systems, human oversight can become thin or nominal (Verdiesen et al. 2021). This raises questions about who, if anyone, can be held morally and legally responsible for wrongful harms (Schulzke, 2013), and whether existing frameworks of international humanitarian law can be meaningfully applied when lethal decisions are partially or wholly automated (Sassoli, 2014). Reduced human involvement is thus seen as a risk not only because it may make manipulation or accidents more likely, but also because it weakens the normative and legal structures that currently constrain how wars are fought.

At the far end of this continuum, there are debates about catastrophic and existential risk from advanced AI. The core worry is that sufficiently capable systems could exhibit goals or behaviors misaligned with human values, pursue those goals strategically, and gain influence over critical systems in ways that humans cannot readily reverse (Zhuang & Hadfield-Menell, 2020). Bostrom's thought experiment, "paperclip maximiser," imagines a system that relentlessly optimizes a narrow objective-like producing paperclips-at the expense of all other values (Bostrom, 2014). More recent discussions focus on the possibility that advanced models might engage in deceptive behavior, resist shutdown, or strategically manipulate human operators (Carroll et al. 2023). Proposed responses include corrigibility, shutdown mechanisms, and alignment research aimed at keeping human oversight meaningful even as AI capabilities increase (Firt, 2025; Hendrycks et al., 2023).

Table 1 summarizes these threat vectors in a two-by-two matrix. It shows how issues that often appear separately in public debate can be located within a single analytic map. Deepfakes and synthetic disinformation fall in the epistemic–misuse cell, while engagement-optimizing recommender systems sit in the epistemic–systemic cell. Autonomous weapons and AI-assisted cyber-physical attacks exemplify kinetic–misuse threats. Meanwhile, the deeper integration of AI into targeting pipelines, critical infrastructure, labor markets, and governance arrangements produces kinetic–systemic risks, including compressed decision cycles and accident-prone systems. Debates about AI safety and catastrophic risk can then be understood as concerns about systemic risks spanning both domains, especially as AI-mediated systems become harder to supervise, audit, and control.

Table 1

Matrix of Threat Vectors

	Misuse (adversarial use of AI)	**Systemic (structural / embedded use of AI)**
Epistemic (informational)	**Deliberate manipulation of information space and beliefs using AI.** *Examples:* ▪ Deepfake video, images, or audio to mislead ▪ Voice cloning to impersonate officials or family ▪ AI-written propaganda and disinformation at scale ▪ AI-assisted phishing and scam messages ▪ Bot and synthetic-persona influence operations ▪ AI-enabled disruption of online communication networks	**Harms arising from how AI is built into platforms and knowledge infrastructures.** *Examples:* ▪ Recommenders that amplify outrage and polarization ▪ Echo chambers and filter bubbles from algorithmic curation ▪ Biased automated moderation over-blocking and under-blocking ▪ Trust erosion as "anything could be fake" ▪ Synthetic content flooding the information ecosystem ▪ Feedback loops where synthetic data contaminates future training
Kinetic (physical / cyber-physical)	**Intentional use of AI to cause or enhance physical or cyber-physical harm.** *Examples:* ▪ AI-assisted attacks on critical infrastructure ▪ Weaponized consumer drones with vision targeting ▪ Unlawful use of autonomous or semi-autonomous weapons ▪ AI support for chemical or biological harm ▪ Sensor spoofing or data poisoning to mislead AI systems	**Harms from integrating AI into infrastructures, industry, and armed forces as part of normal practice.** *Examples:* ▪ AI in targeting and command and control speeding warfare ▪ Arms-race pressures pushing more autonomy ▪ Reliance on AI in logistics, grids, and transport ▪ More "normal accidents" in tightly coupled systems ▪ Cascading failures when widely used models break

2. Deepfake as a Case Study of Epistemic Threat

Epistemic threats operate in the information domain. They aim to shape the information landscape, disrupt communication practices and spread disinformation to undermine trust in societies and institutions. Disinformation is not new. Social

media bots and coordinated campaigns have long disseminated fake news, and disinformation (Hajli et al., 2022). What has changed is the cost and the scale. Recent advances in generative AI reduce the coordination and expertise once required to run large-scale influence campaigns. Generative AI models can craft highly personalized emails or messages in any language, free of the grammatical errors that once signaled fraud (Li and Fung, 2025). Similarly, they can automate phishing and impersonation at volume, and there are now documented cases of AI-generated code and messages being used in credential-theft campaigns (Jabir et al. 2025).

The term "deepfake" emerged in late 2017 and quickly expanded to cover a wider class of synthetic media, generally understood as fake images, audio, or video generated or altered using deep learning models in ways that convincingly depict events or statements that never occurred. In practice, deepfakes include face-swapped videos, AI-cloned voices, fabricated news footage, and fully synthetic personas. Technically, deepfakes draw on advances in generative models such as generative adversarial networks (GANs), which pit a generator model against a discriminator to produce realistic samples (Goodfellow et al. 2020). Subsequent improvements in model architecture, training data, and computation, along with diffusion models and large multimodal architectures, have dramatically improved quality and reduced the expertise needed to produce convincing fakes. As a result, there has been a rapid proliferation. Projections expect over 8m deepfake incidents in 2025 (Khalil, 2025).

Two key features of deepfakes are especially important. First, they have a low marginal cost. Consumer-grade deepfake tools now allow users to generate voice clones from three to five seconds of recorded speech and to create video face swaps or lip-synced clips via user-friendly interfaces (Genelza, 2024). Second, they have high credibility. This combination pushes deepfakes to the center of the AI-driven epistemic threats.

2.1. Threat to Individuals, Society and Political Systems

At the individual level, these threat vectors targeting the epistemic/informational space raise acute questions about rights, dignity, and privacy. The earliest and still dominant use of deepfake technology is non-consensual sexual imagery, including the insertion of a person's face into pornographic content. Empirical analysis has found that the overwhelming majority of such deepfakes are pornographic and disproportionately target women. The abuse of minors with this tool adds a particularly grave dimension, given its overlap with child sexual exploitation material. These practices violate rights to privacy, bodily and sexual autonomy, and often lead to reputational damage, psychological harm, and offline harassment.

Deepfakes distort the notions of identity and representation. They sever the connection between appearance and experience, producing images and voices that are phenomenologically indistinguishable from the person yet bear no relation to

their actual actions or intentions. The generators of these contents have enormous control over one's public persona. They undermine a person's ability to curate how they are perceived, substituting fabricated evidence for authentic self-presentation. From a deontological perspective, using someone's likeness without consent for entertainment, political gain, or extortion treats them as a mere means rather than an end (De Ruiter, 2021).

Beyond individual harms, deepfakes threaten public trust and the epistemic foundations of social life. Regina Rini (2020) argues that the widespread possibility of high-quality synthetic media erodes what she calls the "epistemic backstop" of recordings. Historically, the potential existence of photographic or video evidence could be used for the verification of claims. Yet, as deepfakes spread, both real and fake media become suspect. This produces a dual effect. On one side, citizens may be deceived by fabricated content. On the other hand, genuine evidence may be dismissed as fake by those who find it inconvenient, a phenomenon Chesney and Citron term the "liar's dividend" (Chesney and Citron, 2019). This distortion of epistemic landscape creates a ground in which objective facts are less influential than appeals to emotion and identity. Deepfakes intensify this shift by targeting the very media forms-video and audio-that many people still instinctively regard as trustworthy.

For political systems, deepfakes intensify well-known vulnerabilities. Deepfakes can depict politicians saying or doing things they never did, potentially altering electoral outcomes or inciting unrest (Łabuz and Nehring, 2024). When combined with coordinated distribution through bot networks and micro-targeted advertising, deepfakes can be integrated into broader influence operations that aim to undermine trust in elections, courts, and the integrity of governance. There is also evidence that deepfakes can interact with confirmation bias and motivated reasoning, deepening polarization by providing compelling visual or audio content that appears to confirm prior beliefs (Amin, 2025). Looking forward, the emergence of more agentic AI systems, models with the capability to perceive and act, will only deepen this problem by creating active fake agents which can manipulate the audience with newer information.

2.2. Existing Technical, Legal and Policy Responses

Efforts to respond to AI-enabled disinformation now fall into three broad areas: technical tools, legal rules, and public policy. On the technical side, there is now a large range of detection and provenance tools. These detection methods include deep learning classifiers that identify subtle artifacts in images and videos, analysis of physiological signals such as blinking or heartbeat, and cross-modal consistency checks (Heidari et al., 2024; Mirsky & Lee, 2021; Verdoliva, 2020). A recurring challenge is the gap between offensive and defensive capabilities. While the synthetic

content generators evolve to evade existing detectors, the detection models often lag behind improvements in generation (Lee et al., 2024). Initiatives such as the Content Authenticity Initiative and the Coalition for Content Provenance and Authenticity (C2PA) seek to embed tamper-evident metadata or content credentials into media at the point of capture and editing (Rosenthol, 2022). These standards aim to allow users to verify where an image or video originated and how it was modified. The core mechanism is to trace the content with its metadata to its origin. Large platforms, including TikTok, and major news organizations, have begun to adopt such provenance markers and automatic labeling for AI-generated content.

While authenticity and provenance systems significantly raise the cost of large-scale deception, they are not immune to circumvention. AI-generated content can evade labeling by being produced outside compliant ecosystems, by stripping or laundering provenance metadata, or through re-capture techniques that create new, seemingly authentic recordings. As a result, these systems do not ensure universal detection of synthetic content; rather, they enable verifiable authenticity claims and make the absence or breakage of provenance itself a meaningful signal (Longpre et al., 2024).

On the regulation side, the interest in data provenance is growing. The European Union's AI Act requires that providers of AI systems that generate synthetic content ensure it is appropriately marked as AI-generated and that users are informed when they interact with AI systems, including chatbots and deepfake generators (European Parliament & Council of the European Union, 2024). The Act treats certain manipulative or deceptive AI uses as "unacceptable risk" and thus prohibited, while classifying other uses as "limited risk" subject to transparency obligations. In the United States, multiple legislative initiatives have been introduced in this regard such as DEEPFAKES Accountability Act and TAKE IT DOWN act, while the former failed to pass into law, the latter was successful. TAKE IT DOWN act creates criminal and platform-removal obligations for nonconsensual intimate imagery. Meanwhile, states have moved faster, enacting laws that target specific harms like deepfake pornographic abuse or election-related deception (National Conference of State Legislatures, 2024). However, courts are simultaneously wrestling with First Amendment constraints, as seen in decisions temporarily blocking certain state deepfake laws as overly broad (Waldstricher, 2020). China has adopted a different approach, introducing centralized regulations on "deep synthesis" services that require providers to label synthetic content, prevent malicious uses, and ensure traceability of files (Sheehan, 2023). These rules give regulators significant authority to enforce content moderation and sanction non-compliant platforms. Other jurisdictions, including multilateral bodies such as the United Nations, have begun to frame deepfakes within broader efforts to combat disinformation and protect human rights online (Lederer, 2024).

Policy debates converge on several themes. There is growing agreement on the need for transparency and provenance for synthetic content, protection for individuals harmed by non-consensual deepfakes, and safeguards for democratic processes. At the same time, there is no consensus on the appropriate balance between regulation and free expression, nor on the global interoperability of technical and legal standards. Future policy responses may involve mandatory provenance standards for certain kinds of media, stronger cross-border cooperation on enforcement, and integration of deepfake detection into electoral integrity and national security strategies.

3. Autonomous Weapons: A Kinetic Risk Case

Autonomous Weapons, often discussed as lethal autonomous weapon systems (LAWS), refer to systems that can independently select and engage targets once activated, performing "critical functions" of target selection and attack without further human input (United Nations Office for Disarmament Affairs [UNODA], n.d.). They are distinct from automated or remotely controlled weapons. Remotely piloted systems keep a human operator in continuous control, while automation typically executes pre-set actions without discretionary target choice. Autonomy, by contrast, refers to the system's capacity to sense, decide, and act within predefined parameters without real-time human commands (Truszkowski, 2009). Importantly, autonomy lies on a continuum rather than a binary framework (Sharkey, 2014). The question is, then, whether a system is autonomous, but how autonomy is distributed across the kill chain, and what role humans retain at each stage. Examples of autonomous systems range from semi-autonomous missiles that home on radar signatures, through naval close-in weapon systems that automatically fire on incoming projectiles, to emerging platforms that use machine learning to classify objects and revisit targeting choices as the situation changes. This is why debates focus on forms of human control such as human-in-the-loop, human-on-the-loop, and human-out-of-the-loop, as well as supervisory or shared control models.

Elements of autonomy have existed for decades. Early precision-guided munitions and "fire-and-forget" missiles already embodied forms of autonomy once launched. What is new is the use of computer vision, pattern recognition, and data-fusion algorithms to link detection, identification, tracking, and engagement into a more integrated pipeline (i.e., a full kill chain). In many settings, autonomy is less a single feature than an architectural shift, linking sensors, classification models, and decision-support systems into faster and more tightly coupled chains of action. Contemporary examples include loitering munitions that patrol an area and strike radar emitters with minimal supervision; autonomous air-defense systems that classify and engage projectiles in seconds; and experimental drone swarms that share sensor data and coordinate attacks cooperatively (Atalan et al., 2025) These systems

typically rely on sensors and perception modules to detect and classify objects, navigation and control algorithms, target-selection logic based on rules or learned models, and communications links to other platforms or command-and-control networks.

These developments occur as part of a larger trend in the changing character of warfare (Mattis 2018, 3). Modern militaries and wars are becoming increasingly “algorithmic” (Jensen et al., 2020), with compressed decision times and expanded sensing capabilities. In this concept, military power is increasingly defined by information (Lindsay, 2020). The state that can collect more data, fuse it faster, and exploit it through algorithms gains a relative advantage in identifying threats, closing kill chains, and coordinating forces. With the integration of machine learning systems into command and control, surveillance, reconnaissance, fire-control, and logistics functions, scholars started to describe modern warfare as "agentic warfare” (Jensen and Strohmeyer, 2025). For many modern militaries, autonomy is no longer a question but a perceived necessity due to the vast amount of data volumes, personnel constraints and contested electromagnetic environments that require militaries to adapt to the new reality.

Contemporary conflicts illustrate this trajectory. The proliferation of drones is directly linked to this new concept of military power, and it illustrates how autonomy is reshaping the character of war. In Ukraine, both sides have deployed vast numbers of uncrewed aerial systems-from cheap commercial quadcopters used for reconnaissance and grenade drops to more sophisticated fixed-wing drones and loitering munitions. Official and expert estimates suggest that Ukraine produced around 800,000 drones in 2023, around 2 million in 2024, and aims for up to 5 million in 2025 (Atalan, 2025). Russian forces have likewise launched tens of thousands of drones since 2022, often in large swarms that overwhelm air defenses, with reports of AI-enhanced navigation and jamming resistance (Associated Press, 2025). Swarm tactics increasingly rely on the autonomous movement capabilities of these weapon systems (Grimal and Sundaram, 2018). Although unmanned systems are not necessarily autonomous weapons, this proliferation trend implies that modern wars are increasingly less reliant on warfighters in the front lines and will be more dependent on the autonomous capabilities. Conflict dynamics shape this. These unmanned systems emerged as part of military objectives within the Ukraine case due in large part to Ukraine’s lack of manpower vis-a-vis Russia forces military (Bondar, 2025). First-person-view (FPV) attack drones, guided by small onboard cameras and sometimes partially automated targeting, have become a defining feature of the conflict, enabling cheap precision strikes against armor, artillery, and infantry. As algorithms improve, some systems can autonomously identify specific classes of targets such as vehicles or radar signatures, reducing the need for continuous human piloting. For example, Ukraine delegating target recognition to AI-enabled systems allow locking on to targets up to two km away (Bondar, 2025). This is crucial given

the difficulties of moving frontline personnel through a large drone web. Analysts warn that this increased speed and density of engagements changes the tempo and geometry of battle, potentially favoring those who can deploy swarms of inexpensive autonomous systems.

The motivations differ based on the conflict. In Gaza, investigative reports and expert analyses indicate that Israel has used AI-enabled decision-support systems, including a tool named "Lavender," to process large volumes of intelligence and generate targeting suggestions at scale (Abraham, 2024). Israel's large-scale killings and strategic destruction strategy in Gaza relied on these algorithmic tools. These systems reportedly fuse data from multiple sensors and databases to propose targets for human approval, thereby automating significant parts of the targeting pipeline (Andersin, 2025). The speed and volume of AI-generated target lists raise concerns that human oversight becomes more nominal than substantive (Gusterson, 2024).

3.1. Why it is a threat

Applying the threat vectors framework discussed above, autonomous weapons could pose two vectors of risks. They generate risks through two main sources: misuse and systemic integration. Misuse refers to scenarios in which state or non-state actors intentionally exploit autonomous capabilities for unlawful or destabilizing purposes. Israel's use of AI-enabled decision-support systems to process large volumes of intelligence and generate targeting suggestions at scale is a strong example of this. During Israel's 2-year campaign, more than 15 thousand children in Gaza have died. This number alone raises questions regarding Israel's targeting strategy. On the other hand, this reliance on AI allows adversaries to hack or spoof sensors to redirect autonomous systems, deploy AI-enabled drones deliberately against civilians, or use low-cost loitering munitions for terror attacks. The same features that make autonomous weapons attractive to militaries-lower cost, expendability, reduced need for skilled operators-also make them appealing to actors who seek to cause harm while avoiding attribution and casualties on their own side.

Systemic risk arises from integration into military organizations and routine practice. Delegating life and death decisions to autonomous weapons makes the question regarding responsibility a fundamentally critical issue. Sparrow's "Killer Robots" framed the ethical concern as one of delegating life-and-death decisions to machines and questioned whether responsibility could be meaningfully assigned when no human directly chooses each specific target (Sparrow, 2007). Later debates have distinguished between different levels of human control. Semi-autonomous systems operate with humans "in the loop," authorizing each engagement. Supervised autonomous systems have humans "on the loop," able to monitor and intervene but not approving every action. Fully autonomous systems would, in principle, be "out of the loop," able to select and engage targets without human supervision (Perrin, 2025).

This creates responsibility gaps in practice, where no actor can be clearly held accountable for errors produced by human-machine ensembles. Some argue that as long as a human approves targeting parameters or authorizes system activation, moral responsibility is preserved. Others counter that meaningful human control requires more than a formal veto. It requires that humans understand the system's behavior, have the time and information needed to exercise judgment, and can intervene in practice, not just in theory (Perrin, 2025). If AI systems dominate perception, classification, and recommendation, humans risk becoming mere rubber stamps for machine-generated decisions.

From a legal perspective, the core question is whether autonomous systems can comply with international humanitarian law (IHL), particularly the principles of distinction, proportionality, and precaution. Complex battlefield environments such as urban and irregular conflicts often involve civilians, combatants, and civilian objects in close proximity and proportionality assessments require contextual judgment, including the military value of a target and anticipated collateral damage. Many scholars doubt that current or foreseeable AI systems can make such evaluations in a way that satisfies legal and moral standards, especially when training data encode historical biases or omits rare but critical edge cases (Human Rights Watch [HRW], n.d.). Even if an algorithm performs well on average, humanitarian law is sensitive to catastrophic errors, the small number of failures that can produce large and unlawful harm.

Autonomous weapons may also affect the escalation dynamics. Integration of LLM capabilities into the Command and Control (C2) systems might face a risk of algorithmic drift. Recent benchmark studies show that LLMs are not perfect and show biases (Atalan, 2025). At the end of the day, LLMs are what society makes of them, as they are the results of the training data and model architecture. Different foundational models show significantly different levels of escalation biases for the same crisis scenario. That means the future of C2 systems could inherit these model-level biases. Furthermore, there is also a concern that autonomous weapons can destabilize the escalation dynamics by lowering the threshold for resorting to force (Horowitz, 2021). If states can deploy autonomous systems without risking their own soldiers' lives, political leaders may find it easier to initiate or prolong conflicts. In addition, being unmanned and autonomous makes these systems vulnerable to hacking or spoofing which increases the likelihood that they could be turned against their operators. This is why several accounts argue that AI will create the chance of "flash wars" (Johnson, 2023). Horowitz (2021) emphasizes the destabilizing potential of speed, warning that autonomous systems operating faster than human reaction times can create new escalation risks if misperceptions or errors propagate through tightly coupled systems. Similarly, to the extent that AI/ML alters military power, it could affect how states perceive the balance of power (Horowitz, 2021). As perceptions about power and influence change, it could trigger inadvertent escalation risks

(Johnson, 2023). At the tactical level, the speed of autonomous weapons systems could lead to inadvertent escalation while also undermining signaling commitment during a crisis (Wong et al., 2020). These developments also have arms race dynamics. Autonomous weapons and AI-enabled drones are relatively cheap compared to traditional platforms such as fighter jets or main battle tanks. This cost asymmetry encourages both states and non-state actors to invest in unmanned and increasingly autonomous systems, potentially diffusing advanced military capabilities to a broader set of actors.

3.2. Existing Measures and Responses

Governance efforts for autonomous weapons have unfolded primarily in multilateral forums and civil society campaigns. Since 2014, the Convention on Certain Conventional Weapons (CCW) has hosted a Group of Governmental Experts on LAWS, tasked with exploring the legal, ethical, and security implications of emerging autonomous weapon technologies. While the group has developed guiding principles and acknowledged the need for human responsibility, it has not produced a binding treaty. In parallel, civil-society coalitions such as the Campaign to Stop Killer Robots, led by organizations including Human Rights Watch, advocate for a new international instrument that would prohibit fully autonomous weapons and require meaningful human control over decisions to use force (HRW, n.d.). Opinion polls across multiple countries show majorities opposing weapons that can kill without human oversight, though support can rise when respondents are told that rivals might deploy such systems (HRW, 2021). The scientific community has also shown concerns. In 2015, an open letter on autonomous weapons, organized by the Future of Life Institute and signed by thousands of AI and robotics researchers, warned of a potential military AI arms race and called for restrictions on systems that select and engage targets without human intervention (Future of Life Institute, 2015). More recently, the Responsible AI in the Military Domain (REAIM) initiative has provided a complementary multi-stakeholder forum and produced non-binding outcome documents that offer action-oriented recommendations and practical guidelines for responsible military AI consistent with international law and human accountability (UNODA, 2025).

Despite these efforts, there is currently no specific, universally binding legal framework banning autonomous weapons. States that favor regulation rather than prohibition argue that existing IHL, including requirements for legal reviews of new weapons, is sufficient if properly applied (Nieczypor & Matuszak, 2025). Some governments and alliances have adopted policy guidelines emphasizing responsible AI use, risk assessments, and human oversight in military applications, but these vary in detail and transparency. As autonomous and AI-enabled systems become more embedded in military practice, the window for preventive regulation may narrow.

Looking ahead, the integration of AI into warfare suggests not only a need for diplomatic and legal initiatives, but also for rigorous testing, evaluation, and benchmarking of military AI systems (Jensen and Reynolds, 2025). Robust evaluation frameworks should assess not only technical performance and robustness, but also compliance with IHL principles, escalation behavior in simulated crises, susceptibility to manipulation, and the degree to which human operators can understand and override system behavior. Without such systematic evaluation, militaries risk fielding increasingly opaque and powerful systems faster than they can be governed.

Conclusion

AI poses risks to individuals, societies and political institutions in both epistemic and kinetic domains. It creates risks in the epistemic domain, where harms operate through belief, trust, and the integrity of public knowledge, and in the kinetic domain, where harms involve physical and cyber-physical systems. Across both domains, the sources of risk are not the same. Some harms arise from deliberate misuse by adversarial actors who employ AI to deceive, coerce, or disrupt. Others arise from systemic dependence on AI inside organizations and infrastructures, where decision authority shifts and human oversight can thin out even without malicious intent.

This chapter mapped that landscape and then used two cases to make it concrete. Deepfakes show how AI can undermine the epistemic foundations of public life by making visual and auditory evidence easy to manufacture and easy to contest. The consequence is not only more deception, but weaker verification. As synthetic media becomes cheaper and more convincing, governance pressures shift toward platform accountability, provenance, and standards for authentication that can sustain trust in communication systems. Autonomous weapons and AI-enabled targeting pipelines illustrate the kinetic side of the problem. Even when pursued as a military necessity, autonomy risks decoupling lethal force from robust human judgment. That raises familiar concerns about accountability under international humanitarian law, but it also introduces newer risks tied to speed and opacity, including compressed decision cycles, brittle human oversight, and escalation pathways that become harder to manage in real time.

The two domains are also connected. The same generative models that enable deepfakes can be integrated into influence operations, used to impersonate officials, or deployed to confuse operators in crisis settings. Conversely, AI systems in targeting and command support draw on methods and models that often originate in civilian applications, creating spillovers and dual-use dynamics that are hard to contain. When AI is embedded across social platforms, critical infrastructure, and military organizations, failures in one area can propagate into others. A degraded information environment can worsen crisis signaling and misperception. A rushed targeting pipeline can generate political backlash and feed information operations. These interactions are a central feature of the threat landscape, not an exception.

The future threat landscape is even further complicated. Emerging agentic AI capabilities will likely complicate this picture. The term is often used loosely, but the underlying direction is clear. Systems are moving toward tools that can perceive context, plan, and act across digital environments with less direct supervision. If that trajectory continues, it will not only amplify existing risks in both epistemic and kinetic domains but also change their character. Agents that can persist, adapt, and coordinate across platforms could make manipulation more continuous and more personalized. In security contexts, greater autonomy may further compress decision time and widen the gap between human intention and machine-mediated action. Questions about privacy, responsibility, and trust will become harder to resolve when systems can act at scale while remaining difficult to audit.

Ultimately, the question is not whether AI will transform societies, but how its transformative power is distributed and governed. Responses will have to be layered. Technical measures such as detection tools, provenance standards, and robust evaluation and red-teaming are necessary, but they cannot carry the burden alone. Legal and policy instruments, including emerging transparency regimes and efforts to set limits on autonomous weapons, will need to address hard normative questions about human agency, dignity, and the acceptable distribution of risk. Civil society and public deliberation remain essential, not as a supplement, but as a source of legitimacy and accountability for both states and firms. The choices made about design, deployment, and governance will shape not only economic outcomes, but also human dignity, public trust, and stability in political institutions.

References

Abraham, Y. (2024, April 3). “Lavender”: The AI machine directing Israel’s bombing spree in Gaza. +972 Magazine. https://www.972mag.com/lavender-ai-israeli-army-gaza/

Amin, A., Hong, Y., & Mazhar, B. (2025). The influence of social media deepfake images on political ideology and polarization: The mediating roles of cognitive load and confirmation bias. *Journal of Visual Literacy, 44*(3), 321–339.

Andersin, E. (2025). The use of the “Lavender” in Gaza and the law of targeting: AI decision-support systems and facial recognition technology. *Journal of International Humanitarian Legal Studies, 1*(aop), 1–35.

Associated Press. (n.d.). Swarms of Russian drone attacks Ukraine nightly as Moscow puts new emphasis on the deadly weapon. Retrieved December 15, 2025, from https://apnews.com/article/russia-ukraine-war-drones-putin-attacks-0a06736dea37114aa0f351d21746912b

Atalan, Y., Jensen, B., Reynolds, I., Woo, A., Garcia, M., & Chen, T. (2025). Critical foreign policy decisions (Cfpd)-Benchmark: Measuring diplomatic preferences in large language models.

Atalan, Y. (2025). How Russia is escalating the drone arms race. Foreign Policy. https://foreignpolicy.com/2025/09/23/russia-ukraine-war-drone-missile-poland-nato/

Blauth, T. F., Gstrein, O. J., & Zwitter, A. (2022). Artificial intelligence crime: An overview of malicious use and abuse of AI. *IEEE Access, 10,* 77110–77122.

Bondar, K. (2025). Ukraine’s future vision and current capabilities for waging AI-enabled autonomous warfare. Center for Strategic and International Studies.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford University Press.

Briggs, A., & Burke, P. (2009). A social history of the media: From Gutenberg to the Internet. Polity.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

Brynjolfsson, E., & McAfee, A. (2017). Machine, platform, crowd: Harnessing our digital future. W. W. Norton & Company.

Carroll, M., Chan, A., Ashton, H., & Krueger, D. (2023). Characterizing manipulation from AI systems. In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO ’23). Association for Computing Machinery. https://doi.org/10.1145/3617694.3623226

Chesney, R., & Citron, D. K. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs, 98*(1), 147–155.

DeCanio, S. J. (2016). Robots and humans-complements or substitutes? *Journal of Macroeconomics, 49*, 280–291. https://doi.org/10.1016/j.jmacro.2016.08.003

De Ruiter, A. (2021). The distinct wrong of deepfakes. *Philosophy & Technology, 34*(4), 1311–1332. https://doi.org/10.1007/s13347-021-00459-2

DeSantis, M. F. (2024). Is artificial intelligence a general-purpose technology? [Thesis]. Harvard University.

Eisenstein, E. L. (1979). The printing press as an agent of change. Cambridge University Press.

Ernst, E., Merola, R., & Samaan, D. (2019). Economics of artificial intelligence: Implications for the future of work. *IZA Journal of Labor Policy, 9(*1), 1–35. https://doi.org/10.2478/izajolp-2019-0004

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Executive Office of the President. (2023, October 30). Executive Order 14110: Safe, secure, and trustworthy development and use of artificial intelligence. *Federal Register, 88*(210), 75191–75214. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

Filippucci, F., Gal, P., Jona-Lasinio, C., Leandro, A., & Nicoletti, G. (2024). The impact of artificial intelligence on productivity, distribution and growth: Key mechanisms, initial evidence and policy challenges (OECD Artificial Intelligence Papers No. 15). OECD Publishing.https://doi.org/10.1787/8d900037-en

Firt, E. (2025). Addressing corrigibility in near-future AI systems. *AI and Ethics, 5*(2), 1481–1490.

Future of Life Institute. (2015, July 28). Autonomous weapons: An open letter from AI & robotics researchers [Open letter]. https://www.stopkillerrobots.org/wp-content/uploads/2013/03/FLI_LtrJuly2015.pdf

Genelza, G. G. (2024). A systematic literature review on AI voice cloning generator: A game-changer or a threat? *Journal of Emerging Technologies, 4*(2), 54–61.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139–144. https://doi.org/10.1145/3422622

Grimal, F., & Sundaram, J. (2018). Combat drones: Hives, swarms, and autonomous action? *Journal of Conflict and Security Law, 23*(1), 105–135. https://doi.org/10.1093/jcsl/kry008

Gusterson, H. (2024). It's all Lavender in Gaza. Anthropology Today, 40(6), 1–2.

Hajli, N., Saeed, U., Tajvidi, M., & Shirazi, F. (2022). Social bots and the spread of disinformation in social media: The challenges of artificial intelligence. *British Journal of Management, 33*(3), 1238–1253. https://doi.org/10.1111/1467-8551.12554

Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14*(2), e1520.

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An overview of catastrophic AI risks. arXiv. https://doi.org/10.48550/arXiv.2306.12001

Horowitz, M. C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas national security review, 1*(3), 37-57.

Horowitz, M. C. (2021). When speed kills: Lethal autonomous weapon systems, deterrence and stability. In Emerging technologies and international stability (pp. 144–168). Routledge.

Human Rights Watch. (n.d.). Killer robots. Retrieved December 15, 2025, from https://www.hrw.org/topic/arms/killer-robots

Human Rights Watch. (2021, February 2). Killer robots survey shows opposition remains strong. https://www.hrw.org/news/2021/02/02/killer-robots-survey-shows-opposition-remains-strong

Jabir, R., Le, J., & Nguyen, C. (2025). Phishing attacks in the age of generative artificial intelligence: A systematic review of human factors. *AI, 6*(8), 174. https://doi.org/10.3390/ai6080174

Jensen, B. M., Whyte, C., & Cuomo, S. (2020). Algorithms at war: The promise, peril, and limits of artificial intelligence. *International Studies Review, 22*(3), 526–550. https://doi.org/10.1093/isr/viz025

Jensen, B., Reynolds, I., Atalan, Y., Garcia, M., Woo, A., Chen, A., & Howarth, T. (2025). Critical foreign policy decisions (cfpd)-benchmark: Measuring diplomatic preferences in large language models. arXiv preprint arXiv:2503.06263.

Jensen, B., & Strohmeyer, M. (2025). Agentic warfare and the future of military operations: Rethinking the Napoleonic staff. Center for Strategic and International Studies.

Johnson, J. (2023). AI and the bomb: Nuclear strategy and risk in the digital age. Oxford University Press.

Karnofsky, H. (2016, May 6). Some background on our views regarding advanced artificial intelligence. Open Philanthropy. https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence/

Khalil, M. (2025, September 8). Deepfake statistics 2025: AI fraud data & trends. DeepStrike. https://deepstrike.io/blog/deepfake-statistics-2025

Łabuz, M., & Nehring, C. (2024). On the way to deep fake democracy? Deep fakes in election campaigns in 2023. *European Political Science, 23*(4), 454-473.

Lederer, E. M. (2024, March 21). *UN adopts first resolution on artificial intelligence to encourage opportunities, mitigate risks*. *AP News*. https://apnews.com/article/8079fe83111cced0f0717fdecefffb4d

Lee, H., Lee, C., Farhat, K., Qiu, L., Geluso, S., Kim, A., & Etzioni, O. (2024). The tug-of-war between deepfake generation and detection. arXiv. https://arxiv.org/abs/2407.06174

Li, M. Q., & Fung, B. C. (2025). Security concerns for large language models: A survey. Journal of Information Security and Applications, 95, 104284. https://doi.org/10.1016/j.jisa.2025.104284

Lindsay, J. R. (2020). Information technology and military power. Cornell University Press.

Longpre, S., Mahari, R., Obeng-Marnu, N., Brannon, W., South, T., Gero, K., ... & Kabbara, J. (2024). Data authenticity, consent, & provenance for AI are all broken: What will it take to fix them? arXiv. https://arxiv.org/abs/2404.12691

Mattis, J. (2018). Summary of the 2018 national defense strategy of the United States of America.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54*(1), 1–41.https://doi.org/10.1145/3425780

National Conference of State Legislatures. (2024, November 22). Summary: Deceptive audio or visual media ("Deepfakes") 2024 legislation. https://www.ncsl.org/technology-and-communication/deceptive-audio-or-visual-media-deepfakes-2024-legislation

Nieczypor, K., & Matuszak, S. (2025, October 14). Game of drones: The production and use of Ukrainian battlefield unmanned aerial vehicles. OSW Centre for Eastern Studies. https://www.osw.waw.pl/en/publikacje/osw-commentary/2025-10-14/game-drones-production-and-use-ukrainian-battlefield-unmanned

Perrin, B. (2025, January 24). *Lethal autonomous weapons systems & international law: Growing momentum towards a new international treaty*. *ASIL Insights, 29*(1). https://www.asil.org/insights/volume/29/issue/1

Rini, R. (2020). Deepfakes and the Epistemic Backstop. *Philosopher's Imprint, 20*(24).

Roff, H. M., & Moyes, R. (2016, April). Meaningful human control, artificial intelligence and autonomous weapons. Briefing paper prepared for the Informal Meeting of Experts on Lethal Autonomous Weapon Systems, UN Convention on Certain Conventional Weapons.

Rosenthol, L. (2022, October). C2PA: The world's first industry standard for content provenance. In Applications of Digital Image Processing XLV (Vol. 12226, Article 122260P). SPIE. https://doi.org/10.1117/12.2632021

Sassoli, M. (2014). Autonomous weapons and international humanitarian law: Advantages, open technical questions and legal issues to be clarified. *International Law Studies, 90*(1), 308–340.

Schulzke, M. (2013). Autonomous weapons and distributed responsibility. *Philosophy & Technology, 26*(2), 203–219. https://doi.org/10.1007/s13347-012-0089-0

Schumpeter, J. A. (1934). Theory of economic development. Routledge.

Sharkey, N. (2014). Towards a principle for the human supervisory control of robot weapons. *Politica & Società,* 2, 305–324. https://doi.org/10.4476/77105

Sparrow, R. (2007). Killer robots. *Journal of applied philosophy, 24*(1), 62-77.

Sheehan, M. (2023). China's AI regulations and how they get made. Horizons: Journal of *International Relations and Sustainable Development*, (24), 108–125.

Truszkowski, W., Hallock, H., Rouff, C., Karlin, J., Rash, J., Hinchey, M., & Sterritt, R. (2009). Autonomous and autonomic systems: With applications to NASA intelligent spacecraft operations and exploration systems. Springer Science & Business Media.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018, March 19). Social media, political polarization, and political disinformation: A review of the scientific literature (SSRN Scholarly Paper No. 3144139). SSRN. https://doi.org/10.2139/ssrn.3144139

U.S. Department of Defense. (2023, January 25). DoD Directive 3000.09: Autonomy in weapon systems. https://www.esd.whs.mil/Portals/54/Documents/DD/ issuances/dodd/ 300009p.pdf

United Nations Conference on Trade and Development. (2025). Technology and innovation report 2025: Inclusive artificial intelligence for development. https://unctad.org/publication/technology-and-innovation-report-2025

United Nations Office for Disarmament Affairs. (n.d.). Lethal autonomous weapon systems. Retrieved December 15, 2025, from https://disarmament.unoda.org/en/our-work/emerging-challenges/lethal-autonomous-weapon-systems

United Nations Office for Disarmament Affairs. (n.d.). Responsible by design: A conversation with the Global Commission on Responsible AI in the military domain. Retrieved December 15, 2025, from https://disarmament.unoda.org/en/updates/responsible-design-conversation-global-commission-responsible-ai-military-domain

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2023). AI in drug discovery: A wake-up call. *Drug discovery today, 28*(1), 103410.

van de Poel, I. (2023). AI, control and unintended consequences: The need for meta-values. In Rethinking technology and engineering: Dialogues across disciplines and geographies (pp. 117–129). Springer International Publishing.

Verdiesen, I., Santoni de Sio, F., & Dignum, V. (2021). Accountability and control over autonomous weapon systems: A framework for comprehensive human oversight. *Minds and Machines, 31*(1), 137–163.

Verdoliva, L. (2020). Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing, 1*4(5), 910-932.

Waldstreicher, B. (2021). Deeply fake, deeply disturbing, deeply constitutional: Why the First Amendment likely protects the creation of pornographic deepfakes. *Cardozo Law Review, 42*(2), 729.

Wong, Yuna Huh, John Yurchak, Robert W. Button, Aaron B. Frank, Burgess Laird, Osonde A. Osoba, Randall Steeb, Benjamin N. Harris, and Sebastian Joon Bae, Deterrence in the Age of Thinking Machines. Santa Monica, CA: RAND Corporation, 2020. https://www.rand.org/pubs/research_reports/RR2797.html.

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values, 41*(1), 118–132. https://doi.org/10.1177/0162243915605575

Zhuang, S., & Hadfield-Menell, D. (2020). Consequences of misaligned AI. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H.-T. Lin (Eds.), Advances in Neural Information Processing Systems (Vol. 33, pp. 15763–15773). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/hash/b607ba543ad05417b8507ee86c54fcb7-Abstract.html

About the Author

Dr. Yasir ATALAN
Center for Strategic and International Studies | aatalan[at]csis.org
ORCID: 0009-0003-9219-4833

Yasir Atalan is a deputy director and data fellow in the Futures Lab at the Center for Strategic and International Studies. His research focuses on technology and international security, with particular emphasis on how artificial intelligence is reshaping conflict dynamics. Dr. Atalan examines how AI agents can be responsibly integrated into defense and foreign policy systems by enabling agentic decision-support workflows. Atalan's research has been featured in the New York Times, CNN, Financial Times, BBC, Foreign Policy, Business Insider, IEEE spectrum, War on the Rocks, and Defense One, as well as in peer-reviewed journals. He also serves as a faculty fellow at the Center for Data Science at American University, where he contributes to computational social science initiatives. Previously, he served as the lead replication analyst for the journal Political Analysis at Cambridge University Press. He was a fellow at the Schmidt Futures' International Strategy Forum during 2023-2024. He received his PhD in political science from American University, an MA in Middle Eastern studies from King's College London, and a BS in political science and international relations from Bogazici University