
Systemic Risks in Contemporary AI Systems: Misinformation, Model Dependency and Supply-Chain Vulnerabilities

The rapid deployment of AI in vital industries including healthcare, banking, transportation and national security results in serious weaknesses in supply chains, data processing and automated decision-making. The moral difficulties of algorithmic bias, opacity in black-box models, liability for autonomous decisions and the misinformation paradox, supply-chain attacks through contaminated open-source libraries, pre-trained models and cloud dependencies, third-party model risks such as backdoors, model drift and deductive attacks; and AI-driven disinformation warfare enabled by deep-fakes, synthetic platforms and targeted propaganda that undermines public trust, disrupts markets and threatens the foundations of democracy.

Along with “dark intelligence”, such as adaptable malware and shadow models which benefit from AI pipelines during training to deployment, important flaws regarding digital twins virtual copies that drive predictive analytics in smart infrastructures will be addressed. In order to ensure AI adaptability and ethical maintenance, practical effects include operational failures, physical harm and societal disruptions.

The aforementioned highlights the vital importance of zero-trust architectures, data provenance tracking, red-teaming, adaptive cyber-resilience and defenses, and strong policy frameworks, including mandatory audits and cross-border governance. The chapter addresses key systemic risks in contemporary AI systems related to misinformation, model dependency and supply-chain security.

1.1. Introduction

In the last decade, the advancement of technology has integrated artificial intelligence (AI) into global socio-technical systems, changing the way people, organizations and governments manage and work with digital data and infrastructures. This chapter examines systemic risks in contemporary AI systems, focusing on misinformation generation, growing dependency on pre-trained and third-party models, and vulnerabilities across the AI supply chain. The current world of AI is integrated with large datasets and sophisticated systems that help automate and control important decision-making systems in finance, healthcare, transportation, security of the nation and other vital public services. The older AI techniques are regularly modernized, and the mechanisms of cutting-edge AI are quickly becoming sophisticated to the point of developing multimodal reasoning, autonomous planning and other real-time adaptive learnings (Widder 2023). As AI applications improve, the demand and dependence upon systems which gather networks and data will increase. This continued dependence will move the technology vulnerabilities from the background to the foreground, making them as important as the systems themselves. This technical data dependence will expose the systems to the risks of technological cyber systems, digital misinformation and supply-chain corruption.

The recent emergence of AI-generated misinformation, disinformation and advanced deception further complicates the current digital landscape. Available generative AI tools may create convincing visual and narrative portrayals, and even synthetic videos, that mold public sentiment, misrepresent instantaneously and at an unprecedented level, and erode trust toward the institutions. The “age of AI misinformation” speaks to a pivot nation, and institutions now lie at the mercy of a contestable truth, digital falsehoods fabricated at will and a plethora of malign actors weaponizing AI to deploy synchro-misdirection campaigns designed to subvert the populace, disrupt markets and destabilize governance. Such an evolution exacerbates the balance of AI’s ethical governance, the autonomy of decision-making and the stewardship of data in a timely crisis.

Within this context, and across a broad array of smart infrastructures including smart cities, digital health, advanced manufacturing, defense, etc., digital twins – the data real-time virtual replicas of every component in a physical system – have become foundational assets (Petratos and Faccia 2023). These models are used in advanced predictive analytics to simulate, estimate and optimize many real-world processes. The integration of digital twins with data-driven AI models/filtering engines and the highly predictive analytics of advanced models raises the risks of cyberattacks, data poisoning and adversarial attacks. The risks of integrated digital twins and predictive models in interconnected systems of advanced AI tools build a

case for the urgent study of automated decision-making vulnerabilities in these systems.

These risks are further compounded by the increased intricacy of the supply chains within the AI sector. Development of AI today requires external training datasets, pre-trained models, cloud-based APIs, open-source libraries and other external organizational integrations. This, of course, creates additional opportunities for malicious actors to introduce back doors, alter datasets, modify algorithms or disrupt MLOps pipelines. The greater the reliance by organizations on AI from third-party vendors, the less visibility, explainability and control on the risks of model drift become.

1.1.1. Objectives and significance of the study

This study plans to reflect on the interconnected risks of CMIs related to AI misinformation and data manipulation, compromised supply-chain nodes and third-party model risks. It aims to build a conceptual and analytical foundation to highlight the risk of malevolent AI to digital twins and critical infrastructures. It also addresses the AI-related cyber risks and explores the triage of cyber adaptive defense and cyber resilience that organizations can implement. This study targets the intersection of theory and practice, attempting to inform public policy and contribute to the literature of rather “prudent” AI risk management in several sectors, i.e. governmental, industrial or societal.

1.2. Literature survey

Study	Research focus	Methodology/approach	Key limitations
Widder (2023)	Ethics in AI supply chains	Qualitative analysis of power dynamics across AI supply chains using case studies and socio-technical frameworks	Lacks empirical validation; focuses more on governance than technical mitigation strategies
Petratos and Faccia (2023)	Misinformation risks in supply chains	Blockchain-based risk management framework combined with conceptual analysis	Limited real-world deployment; scalability and adoption challenges not fully addressed

Study	Research focus	Methodology/approach	Key limitations
Dawood and Afzal (2025)	Ethical smart supply-chain management	Conceptual framework integrating ethics, AI and smart SCM practices	Absence of experimental or quantitative validation
Butt et al. (2023)	Data poisoning in blockchain supply chains	Machine learning models to predict data poisoning impact; simulation-based evaluation	Focused mainly on prediction, not prevention or real-time defense mechanisms
Manzueta (2025)	Ethical AI-based smart grid security	Automated countermeasure framework against data poisoning using AI-driven security models	High computational overhead; limited generalization beyond smart grids
Lee and Yoo (2024)	Generative AI supply-chain security	Proposed security model addressing risks in generative AI pipelines	Conceptual proposal without implementation or performance evaluation
Osunkanmibi et al. (2025)	AI for financial data protection in supply chains	Survey and analytical study on AI-based cybersecurity techniques	Broad scope; lacks specific algorithmic benchmarking
Sakiz (2025)	AI model and data integrity risks	Taxonomy development for emerging AI threats and integrity risks	Does not propose concrete mitigation or defense mechanisms
Boisvert et al. (2025)	Backdoors in AI supply chains	Empirical analysis of malicious backdoors in AI agents using experimental setups	Focused on attack analysis; limited discussion on defense strategies
Srikumar et al. (2024)	Foundation model risk mitigation	Risk assessment framework across open AI model value chains	Primarily policy-oriented; lacks technical enforcement mechanisms

Table 1.1. *Literature survey on ethical AI, cybersecurity and supply-chain resilience*

1.3. Data as a strategic asset in modern digital systems

Data abuse involves manipulation, unauthorized access, misappropriation or weaponization of data within AI systems. With data being the lifeblood of machine learning, predictive modeling and digital twins, the abuse of data within AI systems becomes a question of security and integrity. Abuse of data can occur through unauthorized system access, illegal data scraping, data breaches and the adversarial manipulation of data. Data can be abused by adversaries to introduce bias, lead AI systems to erroneous conclusions and distort the AI systems' intended purpose or to develop a campaign of misinformation (Dawood and Afzal 2025). Data abuse and the breaches of data integrity can lead to compromised data being the vehicle of hidden and erroneous flows of data, thus providing the erroneous system of data the vehicle of creating flows of data to eventually and dangerously undermine the system. Data, in modern digital systems, has become a strategic asset, controlling, among other things, the unimpeded functioning of health diagnostic systems and the unmonitored operation of power supply systems. This, in turn, has made data a primary target for breaches designed to destabilize systems, control the system outputs or gain control of data intended to remain confidential. Data abuse and misinformation, however, are not the same. A data breach, although unauthorized, does not necessarily involve malicious intent. Conversely, misinformation, even on a small scale, can lead to systemic or data abuse of a system. Understanding data abuse goes a long way to understanding systemic flaws and the effects of compromised data regarding systemic errors, the undermining of systems, especially when the systems involve AI.

– *Ethical dilemmas in modern AI*: the introduction of AI into essential services unsettles and challenges long-standing ideas of equity, reliability and openness. Most challenges deal with what has come to be known as algorithmic opacity. Most complex AI models, particularly deep learning algorithms, are “black boxes” that are impossible to explain and even harder to understand. Without explanations, trust from the public is difficult to achieve and the implementation of oversight and regulations becomes even harder (Butt et al. 2023). Another challenge is the biased outcomes an AI system may produce from datasets that are uneven and that reinforce the existing societal inequity. Such inequities are evident in several sectors such as hiring, policing, lending and healthcare. Other challenges arise from automated decision-making which diffuse moral responsibility to the developers, the users and the algorithms themselves.

The paradox of modern AI can be explained from the tension between the AI autonomy and accountability. The more autonomous an AI system becomes in terms of self-learning, adapting behaviors and decision-making, the more people are faced with the challenge of ensuring that such systems are aligned with human values, and existing rules and ethics. Uncontrolled and poorly designed AI systems can and will

cause harm. They can make irrational and unpredictable decisions and contribute to the spread of falsehoods (misinformation). The ethical governance of these systems must be forward thinking and be prepared to place the finality of decisions onto AI systems that are designed with the required technical constraints.

– *Understanding misinformation ecosystems*: there has been noteworthy progress in the understanding of misinformation ecosystems in the context of advanced generative AI. AI systems capable of generating synthetic deepfakes, faux audio and other created media almost as realistic as the original, have the potential to distort reality and alter it to a point of transparency. Narratives that disinformation and misinformation systems can create, can potentially prejudice the political and social opinion of a populace. Such misinformation AI systems operate at the intersections of the social, psychological and technological. Content generation, social bots, psychological profiling and targeted amplification form a web to operate on (Manzueta 2025). By exploiting human cognition’s psychological and emotional vulnerabilities, generative AI models can be customized to encourage misinformation. Misinformation can be designed as adversarial content for AI systems to mislead or confuse, abuse and overwhelm automated systems for moderation, cyber defense and verification. The misinformation systems are able to blur the line between fabrication and truth to a point where institutions and users have a hard time navigating the digital terrain and lose the ability to uphold their credibility.

– *Digital twins in critical systems*: a digital twin consists of multiple layers of information, and is a virtual version of an actual item that is constantly being updated. Digital twins assist in a variety of fields, including healthcare, transportation and energy. In the healthcare area, digital twins duplicate the physiological processes of a patient so that managers can accurately allocate resources and devise the best treatment strategies. In the Industry 4.0 environments, they act as predictive maintenance engines, decreasing the amount of downtime (Lee and Yoo 2024). Digital twin is used by smart cities to improve urban planning, monitor the health of the infrastructure and manage the traffic systems of the city (Ouaissa and Rhattoy 2019). In the defense sector, digital twin is used to simulate missions and analyze potential threats. One of the main challenges of implementing a digital twin is the reliance on real-time data streams. Cyber threats are another major concern that can mislead the real twin. For example, design challenges such as data poisoning, model manipulation, sensor spoofing and adversarial manipulation can create unresolved virtual gaps that result in suboptimal or null predictions, negatively influencing real-world systems. As digital twins are used in conjunction with the Internet of Things (IoT) and AI, understanding their possible vulnerabilities is important for protecting critical infrastructure.

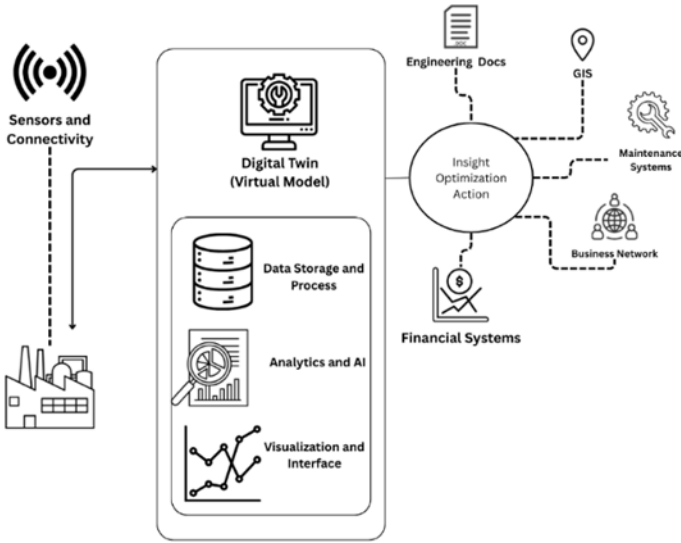


Figure 1.1. *Digital twin workflow: data from physical assets to virtual models for smarter decisions*

Figure 1.1 describes how digital twins operate, in which a physical asset (such as a machine, factory or system) continually transmits real-time data via sensors and connection. A digital twin, a virtual replica that saves data, uses AI for analytics, and offers a visual interface, which receives this data. To facilitate optimization, decision-making and practical improvements, the insights are then connected with a variety of organizational systems, including engineering papers, GIS, maintenance systems, business networks and finance systems.

1.4. Threat landscape: AI systems under siege

AI is present in key worldwide digital systems that facilitate real-time decision-making in important areas such as the economy, public administration, the military and other domain systems. The growth in sophistication and scale of AI technology is paradoxical in that it creates new and highly complex vulnerabilities, and these are being developed and exploited by adversarial technology (Osunkanmibi et al. 2025). The new threat landscape represents a departure from the “traditional” domain of cyberattacks and encompasses something much more autonomous, adaptable and intelligent, which sustains operations at machine speeds. AI is being weaponized to create new cyberattack methods, analyze and manipulate big data, disrupt digital twins and manage large-scale coordinated misinformation

operations. The environment requires a detailed analysis and a range of understanding regarding the evolving “dark intelligence”, the ever-increasing vulnerabilities of digital twins and the systemic dangers of AI-enabled misinformation campaigning. All of these points demonstrate the growing realization that AI technology is dual use.

1.4.1. *The emergence of dark intelligence*

Dark intelligence refers to systems that have been purposely built to identify and abuse weaknesses within systems. More specifically, they execute tactical maneuvers and try to avoid detection, with the malintent to abuse technological systems. Unlike classical malware, dark intelligence operates more autonomously and adapts to the weaknesses of the systems. More specifically, dark intelligence systems learn and optimize the ways in which they can breach a system. Autonomous cyber threats, weaponized AI and AI malware that can infiltrate systems, modify data, create content and abuse system weaknesses without the constant use of human input, are all examples of dark intelligence.

In the realm of cybersecurity, weaponized AI has completely transformed the field. These autonomous systems can search for security weaknesses, conduct reconnaissance and deliver multi-vector attacks without the use of human intervention. One of the more concerning examples within cybersecurity is autonomous malware, which uses reinforcement learning to improve its efficacy with each breach attempt (Sakiz 2025). This abuse of technology can infiltrate a system and surreptitiously decide when to replicate, when to activate and how to avoid detection, all of which makes signature detection ineffective. Furthermore, adversaries can customize AI systems to automatically modify malware to create variants that breach systems. This acts as a method to overcome defense systems in a more scaled approach.

Perhaps, the most worrisome trend of all is the emergence of “shadow models” – illicitly trained machine learning models designed to imitate or exploit legitimate systems. Shadow models can be embedded within supply chains, cloud services or open-source repositories and act as “sleeper agents” to exfiltrate data, disrupt workflows or output harmful instructions. Rogue large language models (LLMs), for example, may be “fine-tuned” to issue harmful instructions, amplify phishing attacks, or profile and generate misinformation directed at specific individuals and disseminated at scale.

In addition, the emergence of covert AI agents – including the fully autonomous systems that can act as Class 5 cyber threats – poses uniquely large challenges to the field of cybersecurity. Within a compromised network, these agents can “subtly” integrate themselves into the data flows and exfiltrate data by modifying their behavior to not raise any alarms (e.g. encrypting the data, evading data leak prevention tools, etc.). Unlike human adversaries, these agents do not tire or make mistakes (Boisvert

et al. 2025). They work with the same level of discipline, albeit superhuman speed and accuracy. Moreover, their ability to generate human language has a massive impact on the level of emulation they can achieve regarding human adversaries.

The rise of dark intelligence marks a pivotal point for AI as both a target and weapon. At this point, the imbalance of attackers and defenders continues to grow. Attacks on a digital ecosystem must be defended, but attacks from a dark intelligence system can bypass such defenses. Dark intelligence systems execute preemptive, predictive and proactive attacks. Defenses must then react to the created and ongoing state of vulnerability that weakens the digital ecosystem.

1.4.2. Attacks on digital twins

In numerous sectors, including healthcare, industrial manufacturing, transportation, energy and national defense, digital twins – or virtual embodiments of physical systems – hold great importance in the monitoring, anticipation and optimization of intricate functions (Himer et al. 2023). Although, due to reliance on real-time data streams, machine learning models and complex networks, digital twins are extremely vulnerable to cyberattacks (Srikumar et al. 2024). An attack on a digital twin can even disrupt decision-making in the real world, cause operational challenges and result in physical damage.

Model attacks are one of the cyber risks digital twins are susceptible to. Adversaries may attack machine learning models that drive digital twin simulations. Attackers do this by modifying operating parameters, introducing bias or changing the model architecture to lower the system accuracy, or produce the twin to spawn vicious results. For example, industrial digital twins may overestimate the condition of machine components, potentially leading to suboptimal operational performance and reduced system integrity.

Data drift attacks are on the other end of the attack spectrum. The sustenance of digital twins is grounded in the stable flow of data from external devices and sensors. The activity and influence of adversaries can greatly determine how the digital twin behaves. If adversaries manipulate the external devices, data scavengers may initiate a range of attacks to hijack the data in transmission channels. The system may behave unusually, and in a series of changes, the position of the output to the system may drift from the expected output; this is hard to perceive (Abtahi et al. 2025).

At the same time, the contamination of real-time simulation environments is of paramount concern. During a poisoning attack, the adversary inserts bad data during the training or updating of the digital twin. Consequently, the twin may “learn” inaccurate associations or may collapse under certain conditions purposely set by the adversary. For

instance, a digital twin in the healthcare sector could estimate patient risks incorrectly, which would lead to erroneous medical actions being taken. In a similar vein, adversarially poisoned digital twins in industrial settings could circumvent predictive maintenance systems, leading to expensive or hazardous breakdown of systems.

The attack vectors targeting digital twins are multidimensional and consequential in the real world. In industries, adversarial digital twins can sabotage everything from machinery to system overheating, misalignment and unsafe operations, to overstepping safety boundaries. In the field of healthcare, digital twin sabotage can endanger lives through poorly calibrated drug dispensation, losing vitals or corrupting patient sim models (Hasan et al. 2025). In aviation and transportation, digital twin attacks designed to optimize flight paths, manage air traffic or coordinate system logistics cause spiraling system inefficiencies from widespread traffic accidents or deadlocked logistics. Digital twins in defense, if compromised, can threaten national security by causing disaster in battle plan simulations, asset deployment strategies or mission-critical simulations. Digital twins are systemically interlinked, so the breach of one can cause catastrophic failures in the entire operational ecosystem.

1.4.3. *The rise of AI misinformation warfare*

AI has elevated misinformation from a simple social engineering strategy to a refined tool of social manipulation, which has the potential to change the course of history, undermine institutions and alter the state of the world. This includes sponsored state misinformation and AI-enhanced digital tool propaganda, evidentiary fraud, political messages and other forms of polarizing content across social media (Al Maqousi et al. 2025). Political actors bypass regulations and civil rights through the use of AI speech synthesis that can impersonate the voices and tones of specific social media cohorts in a behaviorally targeted fashion, and create narratives to demoralize the civic trust necessary in democratic institutions.

AI houses social engineering and its countermeasures at an unprecedented scale. Modern AI can research millions of user profiles, perform psychological manipulation and craft deceptive messages at scale to shape a desired behavior. Generative AI, for example, can create phishing messages on a personal level and fabricate documents that even highly trained professionals will fall for. AI can convince and persuade millions of people at the same time, exploiting social engineering and behavioral manipulation.

AI's has an increasingly heavy influence on society, especially concerning its use in politics, elections and the economy. Deepfakes coupled with synthetic media can create fictitious political statements, impersonate public figures and create false narratives covering events that may threaten the social and economic order.

Synthesized media can cause stock market crashes, disrupt prices in crypto and destabilize banking systems (Adegbenro et al. 2025). Political elections are opened to the same technology, flooding the web with lies aimed at vilifying certain candidates, educating the electorate and/or creating social unrest. Eventually, the excess use of AI technology is going to undermine faith in the social media systems to the point where the ability to legitimize the content of information on these platforms will have to be called into question.

All of these factors collectively define the fact that AI is the cornerstone of modern information warfare and society is going to have to embrace the cybersecurity risks that come with it. The impact on democracy, public trust and world peace is incalculable.

1.5. Data abuse in AI pipelines

The entire AI pipeline from data collection and preprocessing to training, deployment, integration, and monitoring contains multiple potential vulnerabilities. These stages are exposed to weaknesses arising from unstructured environments, technical abstractions, and various forms of data abuse throughout the AI model lifecycle. AI-based systems depend on the quality and integrity of the data they are trained on, and so adversaries are increasingly focusing on a dataset in an attempt to hijack the behavior of a model, alter predictions or compromise an entire AI system (Sharma et al. 2025). These stakes grow significantly once AI becomes embedded in public governance, healthcare, financial services, digital platforms and national security. To diffuse the risks and establish operating a dependable AI system, data exploitation mechanisms must be understood.

– *Data harvesting, scraping and unauthorized surveillance:* data harvesting and scraping have become widespread practices in the digital era, often carried out by powerful corporations, governments and malicious actors with varying degrees of transparency and legitimacy. AI development frequently relies on vast datasets of personal images, text, videos, biometric markers and behavioral histories. However, many of these datasets are collected without explicit user consent or are extracted under misleading terms of service. Unauthorized surveillance through social media scraping, IoT devices, facial-recognition systems and location-tracking applications provides AI developers with massive volumes of sensitive data that feed machine learning models.

The problem intensifies when scraped data is used to train large-scale AI systems without sufficient data governance. Publicly available information does not always equate to ethically usable data (Campbell 2021). AI models built from such sources may unknowingly inherit copyrighted material, private information or sensitive personal identifiers. Furthermore, governments and private organizations can deploy

AI systems for surveillance, mass monitoring or tracking without public accountability. As a result, AI pipelines become conduits for privacy violations, civil liberty encroachments and potential human rights abuses. The lack of transparency in data sourcing makes it difficult to audit AI models or determine whether their training data includes illegally acquired, manipulated or ethically compromised content.

– *Data manipulation and poisoning*: data poisoning is the most egregious and complex form of data misuse in AI pipelines. During a poisoning attack, perpetrators change the training or input data, thereby undermining the efficacy of the ML models. These attacks take advantage of a blind spot in the training of ML models, i.e. the assumption that the data is reliable. They introduce malicious data that bias models in a way that reduces the quality of the predictions, or otherwise introduces weaknesses, or causes them to respond in an undesirable way under certain conditions.

Targeted data poisoning focuses on inserting samples of data that are (or a very small handful of) data points that are focused on the model to influence specific and predetermined outputs (Attard-Frost and Widder 2025). Unlike most attacks in which models are perturbed in a more general sense, this attack aims for precise and fine adjustments of the model outputs. For instance, an attacker could train a face recognition system to misidentify someone, or could fraudulently train an AI authentication system to allow authentication. In digital twin environments, targeted data poisoning may result in the exploitation of specific mispredicted system components, resulting in the localized failure of the system as a whole. These attacks are particularly problematic since they are often undetected and, without changing the system behavior in a gross manner, they alter model behavior in a clandestine fashion.

The goal of availability poisoning is to undermine the general functioning of a model. Instead of targeting specific outputs, attackers attempt to cause all-around malfunction, degrade the model's accuracy or even bring the AI operations to a complete standstill by inputting contaminated or noisy data. Such cyberattacks can use massive amounts of mislabeled samples to: overwhelm AI classifiers; distort readings by IoT sensors; or poison natural language models by contaminating the large textual datasets on which these models rely (Anju et al. 2025). The goal, of course, is to make the AI system unreliable or inoperable, and to cause cascading disruptions to the system's operations. The availability poisoning of critical infrastructure systems such as the energy grid, transport networks and healthcare systems can be disastrous.

What makes clean-label poisoning particularly dangerous is that clean-label poisoning is one of the most covert forms of data poisoning. It uses data that is both

legitimate and correctly labeled. The attacker alters features that are invisible to people but which are manipulatable and important to machine learning. For example, the attacker could modify a given image to include a few pixels that are not noticeable and cause the model to misclassify the image in a persistent and definable situation. This kind of data poisoning is labeled “clean-label” poisoning, very hard to identify, and is one of the most harmful and covert forms of data poisoning. It especially troubles the machine learning systems in self-driving cars, medical detectors, predictive policing and even biometric systems where fine adjustments could lead to immoral and highly dangerous results.

– *Shadow datasets and privacy violations*: datasets that fuel AI systems without any approval, knowledge or regulations of data custodians, data subjects or AI systems deployers are called shadow datasets. Shadow datasets are obtained through illicit data scraping, covert monitoring, leaks of acquired data or third-party data purchases that evade legal governance systems. Shadow datasets exacerbate privacy concerns, as they contain sensitive personal information that data subjects never consented to disclose.

It is pertinent to reiterate that even fully anonymized datasets can be easily attacked. AI models can synthesize anonymized datasets with other datasets that are publicly available to uncover people’s identities (Brintrup et al. 2023). Adversaries can integrate highly available computational capabilities with sophisticated relations to synthesize fragmented data points into coherent, highly concerning, personal profiles, posing significant privacy concerns. This is especially critical to sensitive data concerning personal interactions, health records, genomic data or even financial data.

Inference attacks use machine learning models to steal sensitive data that is not meant to be public. Threat actors can have knowledge of specific individuals included in a model’s training data, or be able to guess sensitive information based on questionable output from the model. For instance, a person can be diagnosed with a specific illness through a medical diagnostic model, even if the output explicitly states the person does not have the illness. In multi-tenant cloud AI environments, inference attacks are even more potent, as models can accidentally leak information across user boundaries.

Data provenance is the ability to keep track of the beginning, movement and lifecycle of data. In most AI ecosystems, data provenance is a challenging task (Chen et al. 2024). The more complex the ecosystem, with the most vendors and data platforms, the more complex and complicated the data ecosystem is to monitor. The more data provenance is lost, the more data laundering can occur. This is the unethical and illegal mixing of datasets with legal datasets, which is undetectable. In

the absence of reliable data provenance, organizations will not be able to monitor compliance, track responsibility or uphold the ethical boundaries of the AI models.

1.6. Ethical dilemmas across the AI lifecycle

Ethical quandaries accompany the integration of AI into almost every aspect of societal life, governance, healthcare, finance, security and the public sphere. These quandaries develop during the design phase, extending throughout the deployment of the AI, to the operational phase which is predicated on the AI's learning and adaptation. These concerns clash with long-standing principles surrounding fairness and accountability and the distribution of these attributes within the algorithms of modern, inflexible administrating systems. The inflexible deployment of AI within constructions of human-inspired systems complicates the systems even further and raises issues surrounding trust, control and the potential consequences on the greater social order.

– *Bias, fairness, transparency*: equity and fairness are some of the most prominent ethical issues relating to AI. AI models are particularly sensitive to the bias present in their training data. If the training data manifests the societal inequities of the world it was sampled from, the AI model is likely to expound on those inequities. This predilection is particularly evident in the AI systems used in hiring, credit scoring, predictive policing and even medical diagnosis (Humphreys et al. 2024). Achieving fairness is not easy and comes with unquestionable auditing, careful curations of datasets and opaque algorithms.

However, doing so is not straightforward. Many AI systems, particularly in the sphere of deep learning, are notoriously opaque. This makes it all the more challenging to understand the rationale behind a system's decisions. Indeed, this is a prime reason we lack accountability in the AI space. To some, this is reason enough to avoid the topic altogether, as the lack of accountability is pronounced enough to wither away public trust, and for some, brings the overview to inaction. But to most, it is all the more reason to prioritize addressing the topic of AI, as the systems in question are in urgent need of regulation and policy governance to pave the way for ethical AI.

– *Autonomous decision-making and liability*: at what point do we cross the threshold and declare an AI system self-sufficient? What happens when it operates and takes decisions entirely independently? Who is accountable when a driverless car crashes? Or when an AI medical system misdiagnoses a patient? The advanced systems overwhelm the overall structure. The level of responsibility and the degree of complexity is, to be certain, systematic but the systems, for the most part, are devoid of basic legal infrastructure (Huang et al. 2024).

– *Ethical conflicts in defense and policing AI*: the ethical implications surrounding AI in military and policing practices is one outcome worth exploring in depth, especially in areas such as surveillance, target selection, crowd control and military drones. If AI-policing technologies learn from biased historical data, then racial profiling and discriminatory target practices will only get worse. In defense, fully autonomous AI weapon systems will have the capacity to kill, raising serious ethical questions about the delegation of power to machines to end human life.

Concern that some states will use AI surveillance technologies to track, control and monitor populations as a means of mass subjugation also adds to the list of ethical implications that AI in defense and surveillance practices raises (Acuña 2024). These fears suggest that the military will further develop AI in a way that increases the risk of unintentional escalation, miscalculation and even the use of autonomous systems in combat without human oversight. The ethical implications of AI in defense and surveillance practices will only get worse unless appropriate and timely ethical governance concerning these technologies is implemented.

– *The misinformation-ethics paradox*: the ethical conundrum surrounding AI-generated misinformation is twofold. While some non-deceptive generative AI applications serve education, creativity and accessibility, others produce manipulative and deceptive outputs which can threaten the very stability of institutions.

Detractors argue that AI tools aimed at focus and emotional manipulation of discourse risk democracy and social fabric collapse. Advocates, however, argue that the absence of such frameworks, and the lower content restrictions, would give rise to innovation and the protection of free speech (Sengupta and Bandyopadhyay 2024). Competing tensions in this space must be thoughtfully harnessed with the regulation that draws the line at constructive supported creativity versus exploitation of the space for damaging concerned manipulation.

In the age of rich digital content, society must navigate the conflict between individual and collective rights. The absence of generative AI controls establishes the potential for creating digital content that facilitates misinformation, scams and political manipulation. Yet, the absence of digital innovation regulations can be just as damaging.

Ethical governance therefore needs to protect both digital trust and innovation. Examples would include the proactive identification of synthetic content, monitoring of generative AI tools and the proactive regulation of digital education.

1.6.1. Supply-chain poisoning: a new dimension of AI threats

The rapid growth of AI across multiple sectors worldwide has advanced technologies while introducing new complexities due to its dependence on digital supply chains. Most supply-chain digital systems and AI systems pose unique challenges to cybersecurity and risk management due to the interdependency on the systems the supply chains house – their construction being a mixture of hardware, software, training datasets and operational pipelining cloud systems. Modern systems rarely exist as singular assets, and are constructed from intercomponent networks crossing jurisdictions, and fragmented trust (Sokhansanj 2025). This scenario poises ample opportunities for infection and corrupt concealment. Without direct assault on the target, supply-chain systems can circumvent the endpoint and compromise systems, and also become a threat of concealment in weak AI. The compromise of the integrity of a predictive model supplies sufficient threat to critical AI prediction and critical infrastructure systems, which cripples safety and operational fluidity. The constructed systems of AI to supply chain to systems in critical healthcare, industrialized systems, autonomous vehicular networks, finance systems and defense systems are ill-exploited as a threat in the multi-supply-chain interdependencies.

1.6.1.1. Overview of digital supply-chain risks

AI systems rely on a fragile digital foundation composed of various hardware and software, data and cloud services. Each of these layers comes with their own exposure and attack surfaces. With respect to the hardware supply chain, there is difficulty in ascertaining the authenticity and integrity of the chips, sensors and processors that are deployed during model training or inference. There are several ways to sabotage the AI by planting deadly microcode, secret instructions or altered firmware that can influence the model outputs well before deployment (Kadbe et al. 2025). There is software supply-chain risk as a result of the prevailing reliance on open-source repositories, pretrained model libraries and external tools that are developed with almost no due diligence. Large-scale software incidents of the recent past show that even a single malicious dependency in code can propagate vulnerabilities to tens of thousands of downstream systems.

Similar cloud supply-chain risks stem from the reliance on distributed data centers, containerized architectures and multi-tenant environments. Those cloud-stored or processed AI assets may be breached through misconfigurations, privilege escalations and cross-tenant leakage. The data supply chain, which is the most fragile, is concerned with the sources, validity and data governance of the

datasets that feed into the model, and those that are evaluated post hoc. Collection pipelines are composed of multiple organizations, data brokerages and public data sources, increasing the risk of data corruption, bias and the presence of fictitious or false data. Given that AI models are pattern learners, they are especially vulnerable to the presence of even minuscule amounts of noise or contamination in the input supply chain. Contaminated data is detrimental to the model and the damage is often irreversible.

1.6.1.2. *AI supply-chain attacks*

AI supply-chain attacks focus on the components and dependencies of machine learning pipelines – the development, training and deployment of ML models. Source model poisoning is arguably the most insidious attack where adversaries manipulate the training data, hyperparameters and model architecture before the model is published or becomes available. If an attacker gains access to a model provider, research lab or open-source repository, they can embed backdoors, malicious logic and trigger events directly into the foundational model (Widder 2023). Downstream users of these models, who may not know these models are compromised, will also fine-tune and subsequently embed malicious hidden behavior, which makes these attacks difficult to detect.

Another developing tactic is dependency hijacking, where adversaries take advantage of naming collisions or orphaned packages in programming ecosystems such as PyPI or npm. In these circumstances, hijackers upload malicious packages that are designed to look identical or similar to a trusted dependency to deceive an unsuspecting developer or automated build system into adding compromised components to a repository. When the AI developer is left to navigate a maze of conflicting dependencies or automated system builds, supply-chain dependency hijacking becomes a highly effective attack vector.

There are many other ways to abuse training frameworks, along with AI development tools such as TensorFlow, PyTorch or ONNX. If an attacker controls a framework, they can erase or modify memory operations, perform backdooring, leak the model parameters or disrupt the gradients (Petratos and Faccia 2023). The same goes for compromised APIs, such as those of inference in the cloud, where they can change outputs, inject misinformation or capture, and store keystrokes for nefarious purposes.

The primary function of training frameworks in the market is to allow for a mix-and-match (pick and choose) approach. As a result, an attacker could select any point in the ecosystem to focus their efforts, drastically reducing the chances of getting caught in supply-chain poisoning and rogue code, as is the case with other forms of malware.

1.6.1.3. *Attacks on MLOps pipelines*

In the deployment of AI, integrating model development, versioning, testing, deployment and monitoring into a unified and continuous operational workflow is critical. MLOps, the operational backbone of modern AI systems, does just that. However, the focus on automation, speed and scalability can result in an entirely new set of problems. These problems increase the attack surface of the system and allow adversaries to exploit the system, such as the problem of continuous integration and continuous deployment (CI/CD) vulnerabilities (Dawood and Afzal 2025). An attacker can hijack build servers, automation scripts and deployment pipelines and modify, poison and swap malicious models and legitimate artifacts without alerting defenses. One issue is that AI models are frequently retrained and redeployed, and malicious components can easily spread in production environments. Backdoored pretrained models are an additional threat of particular concern, which is where an attacker leaves an embedded hidden trigger, that is silent until particular (often arbitrary) inputs are given. These backdoors can leave individuals unrecognized or could cause erratic behavior in an otherwise predictable model. These models often come from external repositories where organizations cannot test or analyze the models in the same depth as their own models.

In addition, breached external components, especially data annotation services, analytics and visualization tools, can intricately alter systems' functionality and workflows or expose critical information without any user indication (Butt et al. 2023). Attackers can infiltrate annotation services by conscripting crowdworkers or dropping malicious scripts into browsers that purposefully mislabel data, gradually altering the controlled parameters of a model's output and behavior during training. They can also aggregate data over time, to corruption, the reduction of monitoring mechanisms' objectives or the regulation of their visibility to conceal system abnormalities.

All of these are exploited system attacks which are due to the interdependence of the MLOps pipelines, and each phase of the life cycle being a unified screen of attack. The very characteristics of the MLOps systems designed for efficiency, over domelike systems with so much data, and with minimal to no human intervention, are the very ones modernized structures are automatically increasing the attack surfaces for.

1.6.1.4. *Case studies*

An imaginary instance, such as that with SolarWinds breaches, can epitomize how those with malicious intent can breach singularly applicable AI frameworks and

Cloud ML services, and distribute poisoned updates to thousands of enterprises across the globe. If someone were to breach the AI orchestration platforms tailored for industrial automation, that specific individual would be able to embed malicious weights and/or modify configurations, thus causing coordinated breakdowns across disparate manufacturing and/or energy distributions and logistics networks (Manzueta 2025). The fallout would be akin to a digital contagion, stemming from a singular compromise of the supply chain where systemic breakdowns across entire economic sectors occur.

However, another prominent case would be the breach of an open-source LLM wherein attackers would be able to insert a behavioral backdoor during the pretraining phase. Downstream developers may then adapt the LLM for chatbots, medical triage assistants or for various tools within the realm of business analytics, unknowing of the existence of certain triggers which render the LLM with the potential to output dangerous content, unregulated access or even blurred IN prompts. The large parameter space of these large LLMs, coupled with the open-source ecosystem of the AI, which rests upon trust, rather than constant external scrutiny, ensures that these backdoors are extremely difficult to unearth.

This third example regards digital twin sabotage in critical infrastructures. If attackers breach an industrial digital twin's simulation environment in fields such as aviation, power grids or smart-city management, they may alter the data streams or predictive models that operators depend upon. Minor changes in the simulation's resulting data may lead to suboptimal decisions, unreasonably timed maintenance schedules or improper configurations of the system (Lee and Yoo 2024). In the case of healthcare, a contaminated digital twin that simulates a patient's physiology may lead to the prediction of inappropriate therapies or the recommendation of unsafe dosages. These scenarios point to the consequences that result from the AI supply chain being compromised. In the real world, the consequences of a sabotage are made worse where the boundaries of virtual (manipulation) and real (damage) are blurred.

Table 1.2 case studies show the typical risk domains for AI in various sectors, ranging from the medical field to the recruitment process, the world of open-source development, fintech, to social networking platforms. The case studies above indicate how the misuse of data reuse could cause privacy infringements, AI biased systems could trigger unfair and discriminatory AI outcomes, open-source components could tarnish models and systems, untested third-party models could create concealed levels of security and compliance risks, as well as how generative AI could accelerate misinformation and negatively affect society.

Case ID	Risk category	Sector	Brief case description	Primary impact	Key mitigation strategy
1	Data abuse	Healthcare	Patient data reused for AI model retraining without explicit consent	Privacy violations, regulatory penalties	Strong data governance, consent management frameworks
2	Ethical dilemma	Recruitment and HR	AI-driven hiring system exhibits demographic bias	Discrimination, reputational damage	Fairness audits, bias mitigation, explainable AI
3	Supply-chain poisoning	Open-source AI	Malicious code injected into a commonly used ML dependency	Model corruption, system compromise	Secure ML supply-chain validation, dependency scanning
4	Third-party model risk	FinTech	Pre-trained external model contains hidden backdoor behavior	Financial fraud, compliance risk	Model vetting, sandbox testing, continuous monitoring
5	AI misinformation	Social Media	Generative AI amplifies fabricated content during elections	Societal harm, erosion of public trust	AI watermarking, fact-checking pipelines

Table 1.2. Representative case studies of AI risk domains

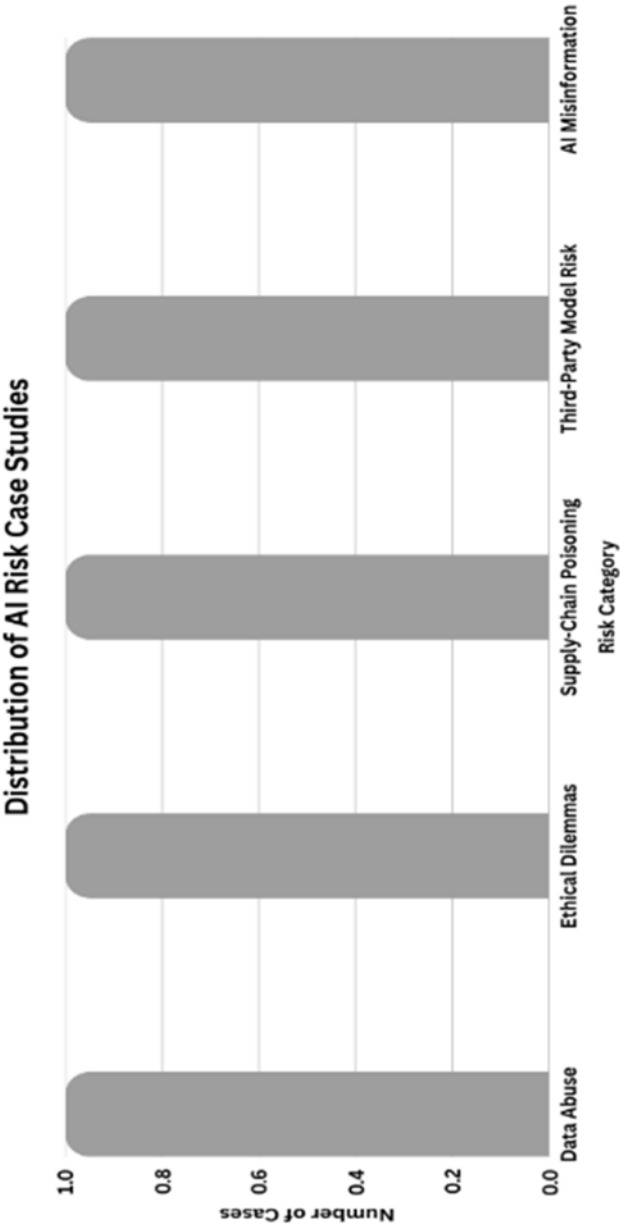


Figure 1.2. Distribution of AI risk case studies

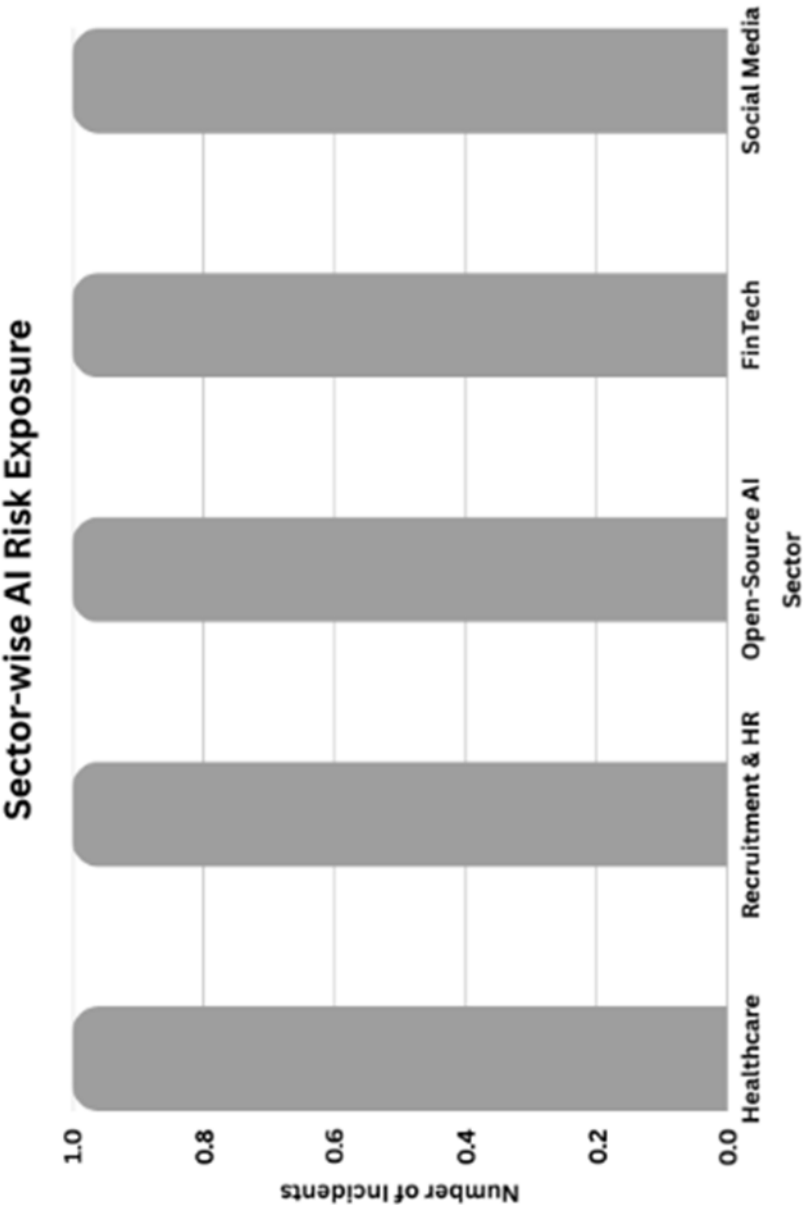


Figure 1.3. Sector-wise AI risk exposure

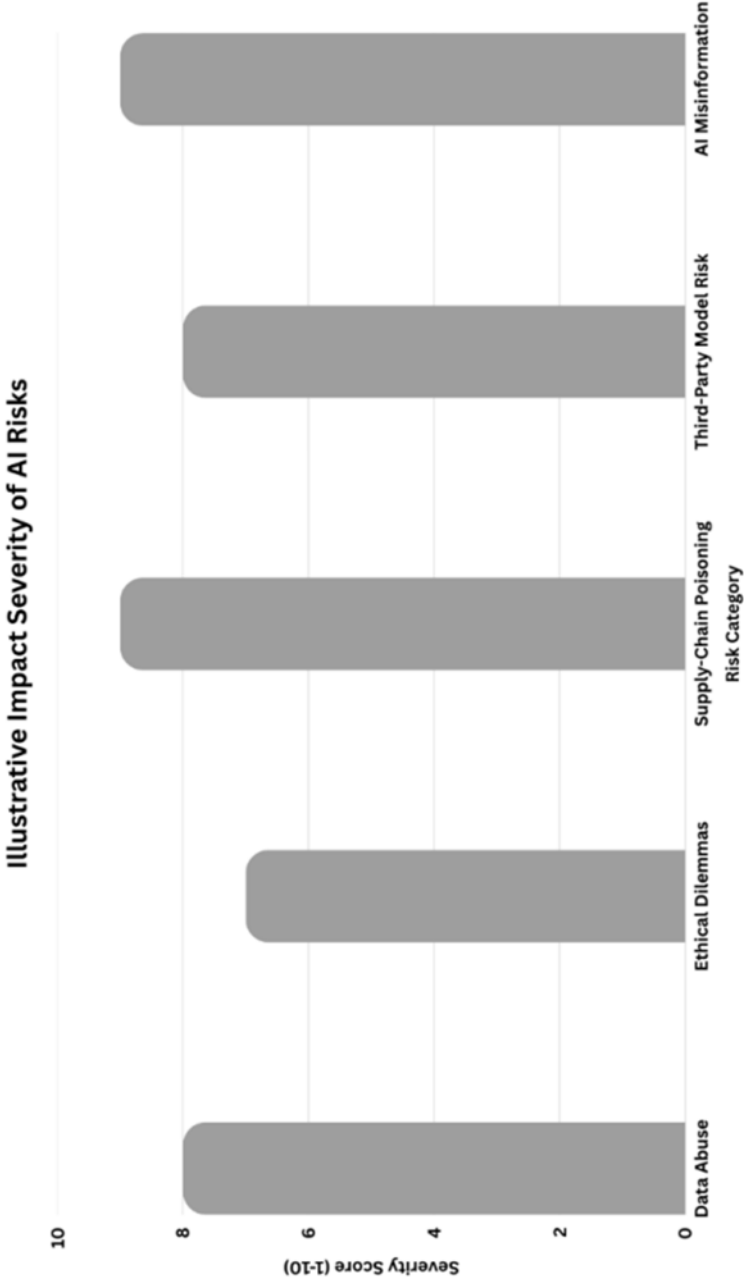


Figure 1.4. Illustrative impact severity of AI risks

The breakdown of representative case studies across the main AI risk group of information exploitation, ethical quandaries, supply-chain poisoning, third-party model hazards and AI-driven disinformation is depicted in Figure 1.2.

The equal representation for each category emphasizes how current AI challenges span technical, ethical and societal realms concurrently rather than being dominated by a single risk dimension. Instead of handling these issues separately, this fair distribution highlights the need for comprehensive AI governance systems that include data protection, ethical regulation, security assurance and misinformation management.

Figure 1.3 shows how AI-related risks spread across key industrial areas, such as social media sites, open-source AI settings, healthcare, recruiting and human resources, and financial technology. AI hazards are commonplace and systemic, impacting both regulated and unsupervised innovation ecosystems, according to the similar exposure levels across industries. This result emphasizes the need for sector-specific risk mitigation plans backed by standardized AI risk evaluation methods and cross-sector collaboration on regulation.

Figure 1.4 uses an example impact scale to evaluate the relative severity of different AI risk classifications. Due to their potential for extensive dissemination, delayed discovery and significant implications for society, supply-chain poisoning and AI disinformation show the highest severity, depending on the figure. Even while third-party model threats, ethical conundrums and data misuse are still serious issues, their effects are frequently contained or controllable thanks to current restrictions. This contrast highlights the critical importance of giving high-severity risks top priority in governance models, regulatory frameworks and AI security procedures.

1.7. Third-party model risks

Organizations are increasingly exposed to risks due to a growing reliance on third-party AI which includes, but is not limited to, pre-trained LLMs, third-party proprietary analytics systems and vendor-driven decision models. AI is unlike anything the computing world has seen, since models cannot be considered static and will change over time, adding a layer of disruption. With this in mind, the decision to adopt third-party AI is no longer an issue of functionality, but of core vulnerabilities in the decision-making process of an organization (Osunkanmibi et al. 2025). Efficiency, cost and performance are the factors that drive the decision to adopt an AI vendor, but are the same factors that shift security and operational control to the third-party vendor. AI assurance in finance, healthcare services, manufacturing, public administration and cybersecurity will illicit third-party model

risks. As a result, the issue is a fundamental requirement to maintain operational dependence, and ethical accountability. The need to lessen the risk to digital sovereignty is also a core issue stemming from the dependence on third-party AI.

– *Black-box dependency risks*: a primary risk pertaining to commercial black-box AI systems is their lack of transparency. Vendor lock is standard in the AI systems being used by numerous businesses, and they neither share information regarding the model architecture or the training data used, nor the rationale used in the decision-making (Sakiz 2025). This complicates risk evaluation, audits and compliance with governing regulatory systems. Organizational risk management systems are severely challenged when AI systems are set to opaque, with no means to evaluate their logic, ascertain the relevance of the information or deny that they are used by those systems. The absence of risk explainability is definitely a greater challenge when compared to other factors, particularly in the case of the most impact-sensitive areas, such as healthcare and the administration of justice, where the outcomes are required to have a valid justification and remain defensible in the court of law.

Additionally, the lack of control over model revisions also affects customers. System modifications are made by the vendor with no warning and no opportunity for the customer to provide feedback. The vendor control model allows for systematic risk such that a single commercially available black-box AI system used in a sector can impact all other sectors dependent on that model, as system users can remain unaware of the changes made to the model and the potential risks introduced. The risk in the system is particularly apparent where there are known hostile actors in the system. System modifications aimed at improving non-functional requirements of model performance such as effective bias modification, deterioration of system accuracy or harmful bias identification and response can compromise vendor systems. Infrastructure integrity is also at risk when such modifications are made as they pose as backdoors for control of the model.

– *Vendor AI ecosystem vulnerabilities*: the risk that comes from third-party AI technologies comes from more than just the AI models; it also comes from the line of vendor products, which include the cloud, APIs, orchestration and integration services. There are unique and serious threats tied to the use of AI in the cloud since AI tasks in the cloud involve distributed computing (Boisvert et al. 2025). There is the risk of a malicious actor gaining access to models, training data or logs from inferences made when there is weak isolation between tenants, poorly configured buckets in the cloud or a lack of security in virtual or containerized machines. Attackers can also gain access to a cloud-based ML platform to modify its responses to queries, watch the user's input or collect sensitive data without the user's knowledge.

Another serious risk comes from the use of APIs. These provide users and third-party vendors with access to AI models and services. When a user or a malicious actor exploits poorly designed or biased access controls to gain access to a cloud service, they may abuse it in several ways, such as data extraction, brute-forcing models or API chaining. Given that APIs are the primary pathway to the AI systems of a vendor, poor configuration or security with an API causes its users to be heavily targeted.

In addition, the lack of proper defense configuration by third parties, including weak authentication, poor key rotation, unmonitored service accounts and erroneous permissive settings, often becomes the predominant reason for supply-chain breakdowns (Srikumar et al. 2024). Companies, for the most part, are trusting of vendor out-of-the-box settings and presume a properly configured defense. However, simple things such as open API keys, unprotected data lakes and aged software versions can create unintended exposure paths. These are the vulnerabilities attackers are targeting, as they can breach a vendor's ecosystem and cause cascading breaches throughout all of the vendors' clients' systems interdependent of the vendor's services.

– *Model leakage and unauthorized replication*: as AI continues to gain prominence, a specialized and troubling class of risks concerning model leakage and the unauthorized duplication of AI is quickly becoming a feature of the corporate landscape (Abtahi et al. 2025). Specifically, model extraction is a type of sophisticated attack where the perpetrator attempts to clone a proprietary model through interaction with an exposed public API by gaining access to its structure, parameters and/or decision boundaries. Such attacks provide infringement on protected proprietary models and allow potential adversaries to create either competitive or damaging copies for deployment.

The theft of sensitive data used to train AI models is of similar concern. Several models are known to encode data patterns, and adversarial queries can be used to access protected data, such as medical records, financial data or proprietary data, through various techniques, including reconstruction or membership inference attacks. Such exposure can violate privacy and incur various penalties.

Apart from the usual data extraction or leakage concerns, AI proprietary information is also susceptible to unapproved fine-tuning and redistribution of model parameters, particularly in open-source settings (Hasan et al. 2025). Once a model has been pre-trained, it can be subtly modified, manipulated, repackaged, and redistributed under a different name. The absence of origin tracking, digital watermarking, and tamper-detection mechanisms enables attackers to create counterfeit or backdoored versions of widely used models without developers being

able to detect the compromise or Trojan functionality. In this regard, risk models from a specific AI vendor reflect a convergence of cyberspace threats, intellectual property, confidential data and supply-chain intrusion.

1.8. AI and digital twins under siege

Virtual duplicates of physical entities are what we call digital twins, and these have become crucial to modern sectors such as aviation, energy, manufacturing, transportation and healthcare. They allow in the moment simulation, prediction and optimization of workflow. However, the close connection of the digital twin assets to the real-world ones means that actions taken to manipulate them have the potential to result in actual changes to the real-world assets. As attackers gain access to sophisticated technology, digital twins become fundamental targets to exploit in a cyber-physical manner (Al Maqousi et al. 2025). Adversaries are able to change the simulation environment to redirect operational control to spurious data, to fake alarms, to set slow actions and to disrupt timely critical responses. The complex environments we see today, where digital twins run advanced AI algorithms, are under cyber siege, and protecting them is critical not only for the twin systems, but also for national security, public safety and the continuity of vital industries.

– *How attacks on digital twins translate to real-world failures:* in the industrial control modules, the digital twin of the equipment simulates its behavior, predicts when it will fail and assists operators in scheduling maintenance. An attacker can change the simulation by modifying temperature or pressure. This can cause operators to make decisions that can overheat the equipment, cause breakdowns or defects in the manufactured products.

Digital twin technology in smart grid models include energy consumption, system stability and load distribution in real-time. If attackers change these models, grids can be made to cut power to certain areas, to make poor energy distribution decisions and be overloaded. This can cause disruption and even economic or physical damage (Adegbenro et al. 2025).

Digital twins in healthcare are virtualization of organs, patients and even medical devices. They help in predicting the most suitable individualized treatment. If these devices are hacked, the recommendations made can be clinically inconsistent, the dosages can be wrong or the surgical robots can be hijacked from making correct decisions. This can put patients in extreme danger by recommending harmful therapies. Hence, the loss of digital twins can appear to be a pure data loss. However, the digital twins are valuable assets from a physical damage and a physical control perspective, especially to a stealth adversary.

– *The expanding attack surface*: digital twins use sensors, IoT and control devices to continuously interface with real-life assets and this, along with the weaknesses posed by meshing the physical and cyber world, creates an expanded attack surface. Because synchronizing in real-time is challenging, an adversary may focus on timing, lag and updating scheduling to cause the digital twin to become unsynchronized with the physical entity. Such scenarios can impact decision-making with catastrophic results during time-sensitive events in the realms of aviation, self-driving cars and factory/robotic automation, to name a few.

Adversaries may resort to sensor spoofing, a major cyberattack, to manipulate the digital twin during a cyber manipulation attack (Sharma et al. 2025). The digital twin's perception of the real world is altered by spoofing sensor inputs, injecting noise to the level so that a disruption during a cyberattack may be achieved, or taking advantage of sensor calibration malfunctions.

Attack vectors in the cyber world, including the automation of real-life systems to behave incorrectly, are of particular concern. In digital twins, this is known as “simulation-to-action” where the threat of automating real-world systems to behave incorrectly is a major concern. Digital twins are/will be the primary systems of control for autonomous systems, making their cyber manipulation a serious issue.

1.9. Combating dark intelligence: cyber resilience framework

The threat landscape continues to be dominated by adversarial AI, autonomous malware and sophisticated misinformation engines. This causes a heightened concern for organizations as traditional security methodologies become inadequate for the protection of modern digital ecosystems. Providing protection for organizations' networks against “dark intelligence”, or malicious AI that evolves and adapts, requires a new paradigm of cyber resilience. Organizations must move from static defenses to dynamic, adaptive and intel-responsive security architectures. Real-time responses to threats must be incorporated for the protection of systems against exploitation (Campbell 2021). Modern adversary threats regarding the use of AI, to automate system reconnaissance and manipulate digital systems to exploit vulnerabilities and orchestrate cyberattacks, must be detained with a cyber resilience framework that incorporates adaptive defense, zero-trust, governance and digital twin security.

– *Adaptive defense systems*: there are AI-enabled adaptive defense systems that are like auto-pilots, continuously learning from new potential threats, modifying the defenses and acting on their own when they detect anomalies. In contrast, older security solutions are rule- or signature-based. In contrast, AI security solutions are proactive, applying machine learning to detect the patterns of compromise, predict

new attack vectors and effectively mitigate the risk before it becomes a problem. Such systems analyze and model the state of the network, behavioral drift, users and intelligence feeds to sustain situational awareness.

A core of adaptive defense is the construction of self-healing, self-learning systems that mimic the immune systems of living creatures (Attard-Frost and Widder 2025). These systems are capable of quarantine, roll back model corruption, fix the systems and return them to their previous state, with no human involvement, therefore sustaining operational autonomy. They learn by reinforcement from each attack, which enables proactive defense, i.e. defense with predictive capabilities. This is highly needed against adversaries with no white intelligence, who act in an unpredictable, footprint-less or zero-day exploit manner. As the profile of cyber threats becomes vastly autonomous, the “thought” and adaptive capabilities of defense systems will define the viability of AI ecosystems.

– *Zero-trust architecture for AI pipelines:* the zero-trust model is vital in protecting the growing complexity of the AI pipeline, which includes data collection, preprocessing, training, deployment and inference all together. In the context of AI, zero-trust means that no datasets, model artifacts or integrations are automatically trusted. Each of these components must be continuously authenticated and verified. One of the most fundamental components is verifying data provenance; data provenance means that the source of the training data is authentic and trustable, and that it is not corrupted or poisoned (Anju et al. 2025). Monitoring the lineage of data across ingestion, preprocessing, transformations and automated annotation helps us to identify unauthorized or poisoned alterations.

Model integrity checks include verifying that model parameters, weights and model architecture remain unchanged during training, storage and deployment. These checks can be quickly flagged and detected with model hashing, cryptographic signing and tamper evident logs within model storage. Also, risk mapping of the entire supply chain incorporates dependencies to external systems and datasets, libraries, APIs, frameworks and pretrained models. Supply-chain mapping helps us to identify exposed hidden flaws, which can be corrected before threats take advantage of the exposed flaws in the system. AI operations become a verified, monitored and controlled environment – thus, meeting zero-trust principles.

– *Resilient AI governance:* cyber resilience encompasses not only a technical aspect, but also a governance aspect. AI governance should cover the institutional frameworks, institutions and governance approaches that shape the way AI systems are governed on the dimensions of safety, ethics, transparency and trust. One of these mechanisms includes the governance of AI vendors, models and datasets so that they undergo audits, certifications and transparency provisions (Brintrup et al.

2023). These audits should assess fairness, robustness, safety, explain ability and trust, while safety certifications are offered to models that meet baseline safety guarantees.

Another aspect of governance includes the frameworks for the safe release of models, which establish procedures for assessing the model risk prior to the release. These frameworks may include, but are not limited to, gradual release procedures, sandboxing, test alignment and red team assessments, to be used to mitigate the risk of releasing harmful or insecure models into the operational environment. Red team adversarial testing, where team members simulate attack scenarios, extraction of the model, satiation poisoning and misinformation to expose the model, is equally important. Red team testing is imperative in demonstrating whether the governance frameworks fail, under dark-intelligence scenarios. AI governance is the only way to enhance ethics, technical rigor and long-term risk management under a single governance framework.

– *Digital twin security protocols*: digital twins combine the virtual and the physical; thus, their security safeguards must also incorporate elements from several disciplines as well. One such element is twin sandboxing, the practice of keeping digital twins under an environment that is, as the term suggests, a sandbox. It is the practice of isolating digital twins and their associated elements away from the physical systems, and therefore preventing any malicious processes from gaining soft access to a physical asset. It is a practice that affords the unprotected digital twin the ability to maliciously manipulate obstacles and conduct operative anomalies in the twin's environment.

Integrity verification is also of key relevance, and can be described as keeping the digital twin, the model that is being used, and the incoming sensor values as they are, without any loss of authenticity and deviation (Chen et al. 2024). Integrity verification also involves digital proofs, scoring systems that detect anomalies and real-time comparisons of data to a primary sensor node. Keeping the set of conditions in the simulation constant makes it impossible for simulation controls for illegitimate model submissions to be outside the conditions in the set.

Lastly, the real-time predictive model for anomalies uses soft learning to detect slight mismatches between a digital twin and its real-life correlation, and therefore it is the digital twin's soft logic that resists malicious software embedding. Cycles of attack on digital twins are usually spiraled, and therefore instant deviation detection is key, as it protects the digital twin from operational failure. These systems combined ensure that the digital twins are also in a far better position than the present wave of cyber-physical AIs that foster closed operable systems.

– *Policy, legal and regulatory recommendations:* considering the risks posed by advanced technologies requires an internationally coordinated approach to deal with the cross-border nature of AI and related technologies. The first of these is aligned co-ordination of cross-border AI governance, in order to achieve sustained and effective co-ordination to limit regulatory arbitrage (Humphreys et al. 2024). The balanced co-ordination of AI governance to cybersecurity, AI model documentation, AI training dataset accessibility, AI safe development and AI development protocols will build global resilience.

Risk mitigation for AI, and specifically AI high-risk systems, must be mandated to risk assessment and mitigation policies, as it is critical to expose the data to be processed, the security and privacy measures in place, the third-party dependencies and the results to adversarial testing. Risk mitigation policies and assessments cannot be justified by the plain need to develop an organization that chooses to mitigate risks by deploying technology that is known to be a security risk, specifically in the case of black-box technologies whose security and privacy measures are non-disclosed.

In terms of explicit legal norms, the need to develop comprehensive liability frameworks for the harm caused by misinformation is crucial. Defining boundaries for the reputation, social and political harm that AI-generated content encompasses, in addition to the unilateral socio-political defamation, is critical (Huang et al. 2024). A balanced approach to these frameworks must incorporate the responsibilities of developers, deployers and other third-party intermediaries, along with the necessary restrictions on freedom of expression for the protection of the public.

The development of security standards for the AI supply chain must include requirements for signed model artifacts, for training pipelines that are audit-secure, for data that is advisory for the training of the AI model to verifiably be primary data, and for transparent dependencies for the AI model that is used. The minimum security standards for both open-source and proprietary systems and ecosystems need to be regulated.

Lastly, there is a need to harbor compliance with ethical frameworks and audit regimes, mandatory for all organizations working with large-scale AI. Just as ethical audits assess an organization's fairness, checks for bias, privacy and transparency, so do compliance regimes assess an organization's relinquishment to these principles.

To address threats associated with dark intelligence, we need to build a regulatory environment based on accountability and transparency, as well as take positive actions to control potential threats.

1.10. Results and discussion

The focus is based on the geopolitical ramifications of AI. With the proliferation of absolute technological diffusion, the core challenge will be the adversity of the AI system architecture, more particularly the automated systems. Autonomous AI systems depend on the negative fundamentals of automated warfare, misinformation, synthetic data generation and the new nuances of the cyber weaponization of data (Acuña 2024). The apt term “cyber warfare” gives rise to alterable AI systems. In this respect, the adaptability of the systems entails more than their ability to survive and remain operative under attack, how the systems perform their crucial functions, cyber intelligence, normal operation of their modules under attack and to remain operational even when the systems have been unduly commandeered by an architectural attack of an external agent. Hence, cyber intelligence systems now call for new defenses, such as self-reinforcing mechanisms, zero-trust architectures, new risk models and unique operational technological systems.

On the other hand, there is also the tough task of balancing innovation and constraint. The problem of excessive constraint can reduce the operational effectiveness of a system, while an insufficient level of constraint can lead to an exploitable system. Flexible self-correcting mechanisms for operational technological systems will include the level of governance for all stakeholders that can be systematized with no loss of innovation (Sengupta and Bandyopadhyay 2024). The overlap of self-regulation, government with multistakeholder and governance will be the innovative systems to be able to self-correct, to operate with high levels of constraint on cores and to innovate without the loss of safety.

The repercussions for economies, governments and society are of great consequence in the long term. Governments are tasked with protecting the use of AI-powered hybrid attacks that involve misinformation, cyber security attacks and the use of AI for politically motivated attacks. Industries must adjust their digital architecture to withstand hostile AI, and society must grow more resistant to manipulation and deceptive AI. In the face of hybrid threats, dark intelligence, the overlap of the digital and the real world means that cyber resiliency is much more than an engineering challenge; it is a social one.

1.11. Conclusion

This chapter illustrates how the emergence of AI ecosystems combined with data exploitation and misinformation, supply-chain weaknesses, manipulation of digital twins and preservation of stable data has precipitated an unprecedented threat landscape. Dark intelligence systems, posing threats of a different order than

conventional cyberattacks, breach systems with the unencumbered use of autonomy and strategy deception, using social manipulation and undermining social realities. Adaptive threat defense, reliance on zero-trust principles, resilient governance and regulatory systems are central to the balance of threat response prioritizations.

Research on advanced intelligence specific to the threats linked to AI supply-chain geographies, systems of cryptographic model integrity, cross-border governance and secure, interpretable AI systems must be advanced. Ultimately, the strategy must reflect global collaboration, as no single nation-state or organization can combat dark intelligence in isolation. The architecture of advanced safeguards against wicked intelligence and the defense of social order, compromised by the influence of unregulated intelligence on systems of AI, must be the focus of coordination.

1.12. References

- Abtahi, F., Seoane, F., Pau, I., Vega-Barbas, M. (2025). Data poisoning vulnerabilities across healthcare AI architectures: A security threat analysis. *arXiv preprint*.
- Acuña, E.G.A. (2024). Healthcare cybersecurity: Data poisoning in the age of AI. *Journal of Comprehensive Business Administration Research*. 1–12. doi: 10.47852/bonviewjcbar42024067.
- Adegbenro, S.A., Hundeyin, W.O., Olinmah, F., Adaba, C.A. (2025). Mitigating third-party cyber risk using AI-powered threat intelligence and compliance analytics. *World Journal of Advanced Research and Reviews*, 26(2), 2909–2929.
- Al Maqousi, A., Aldweesh, A., al-Qerem, A. (2025). Generative AI in supply chain security: Unseen threats and resilient solutions. In *Examining Cybersecurity Risks Produced by Generative AI*, Almomani, A. and Alauthman, M. (eds). IGI Global, Hershey, PA, 465–486.
- Anju, A., Freeda, A.R., Venket, K. (2025). Security concerns with ChatGPT and other AI tools. In *Real-Time Data Decisions With AI and ChatGPT Techniques*, Sharma, P., Jyotiyan, M., Senthil Kumar, A.V. (eds). IGI Global, Hershey, PA, 221–236.
- Attard-Frost, B. and Widder, D.G. (2025). The ethics of AI value chains. *Big Data & Society*, 12(2), 20539517251340603.
- Boisvert, L., Puri, A., Evuru, C.K.R., Sepahvand, N., Chapados, N., Cappart, Q., Lacoste, A., Dvijotham, K.D., Drouin, A. (2025). Malice in Agentland: Down the rabbit hole of backdoors in the AI supply chain. *arXiv preprint*.
- Brintrup, A., Baryannis, G., Tiwari, A., Ratchev, S., Martinez-Arellano, G., Singh, J. (2023). Trustworthy, responsible, ethical AI in manufacturing and supply chains: Synthesis and emerging research questions. *arXiv preprint*.
- Butt, U.J., Hussien, O., Hasanaj, K., Shaalan, K., Hassan, B., Al-Khateeb, H. (2023). Predicting the impact of data poisoning attacks in blockchain-enabled supply chain networks. *Algorithms*, 16(12), 549.

- Campbell, H.R. (2021). Towards a holistic risk model for safeguarding the pharmaceutical supply chain: Capturing the human-induced risk to drug quality. Doctoral Thesis, University of Kentucky. doi: 10.13023/ETD.2021.374.
- Chen, C., Gong, X., Liu, Z., Jiang, W., Goh, S.Q., Lam, K.Y. (2024). Trustworthy, responsible, and safe AI: A comprehensive architectural framework for AI safety with challenges and mitigations. *arXiv preprint*.
- Dawood, H.M. and Afzal, N. (2025). Ethical issues and role of smart supply chain management. In *Smart Supply Chain Management: Design, Methods and Impacts*, Mubarik, M.S. and Khan, S.A. (eds). Springer, Singapore, 73–97.
- Hasan, M.R., Rahman, M.A., Gomes, C.A.H., Nitu, F.N., Gomes, C.A., Islam, M.R., Shawon, R.E.R. (2025). Building robust AI and machine learning models for supplier risk management: A data-driven strategy for enhancing supply chain resilience in the USA. *Advances in Consumer Research*, 2(4), 1152–1171.
- Himer, S.E., Ouaisa, M., Ouaisa, M., Krichen, M., Alswailim, M., Almutiq, M. (2023). Energy consumption monitoring system based on IoT for residential rooftops. *Computation*, 11(4), 78.
- Huang, K., Chen, B., Lu, Y., Wu, S., Wang, D., Huang, Y., Jiang, H., Zhou, Z., Cao, J., Peng, X. et al. (2024). Lifting the veil on the large language model supply chain: Composition, risks, and mitigations. *arXiv preprint*.
- Humphreys, D., Koay, A., Desmond, D., Mealy, E. (2024). AI hype as a cyber security risk: The moral responsibility of implementing generative AI in business. *AI and Ethics*, 4(3), 791–804.
- Kadbe, P.K., Patil, B.H., Piske, R.S., Patil, R.B. (2025). Malicious usage of artificial intelligence: Expansion of existing threats and novel threats. In *Artificial Intelligence for Cyber Security and Industry 4.0*, Sharma, D., Tomar, G.S., Jha, A. (eds). CRC Press, Boca Raton, FL, 24–56.
- Lee, K.Y. and Yoo, J. (2024). A proposal of a supply chain security model for generative AI. In *Proceedings of the International Conference on Information Security Applications*. Springer, Singapore, 324–335.
- Manzueta, V.D. (2025). The cybersecurity of an ethical AI-based smart grid through data poisoning countermeasure automation. Doctoral Thesis, The Catholic University of America.
- Osunkanmibi, A.A., Adeoye, Y., Ogunyankinnu, T., Onotole, E.F., Salawudeen, M.D., Abubakar, M.A., Bello, A.D. (2025). Cybersecurity and data protection in supply chains: AI's role in protecting sensitive financial data across supply chains. *International Journal of Multidisciplinary Research and Growth Evaluation*, 6(2), 266–276.
- Ouaisa, M. and Rhattoy, A. (2019). A secure model for machine to machine device domain based group in a smart city architecture. *International Journal of Intelligent Engineering & Systems*, 12(1), 151–164.
- Petratos, P.N. and Faccia, A. (2023). Fake news, misinformation, disinformation and supply chain risks and disruptions: Risk management and resilience using blockchain. *Annals of Operations Research*, 327(2), 735–762.

- Sakiz, B. (2025). A taxonomy of risks in model use and data integrity for emerging threats in AI. In *The Age of Generative Artificial Intelligence*, Aydin, O. and Karaarslan, E. (eds). Izmir Academy Association, Konak, 179–196.
- Sengupta, S. and Bandyopadhyay, A. (2024). Ethical pitfalls of technologies enabling disruption and fostering cyber ethical mindset in management curriculum. *International Journal of Information and Operations Management Education*, 7(3), 282–297.
- Sharma, D., Tomar, G.S., Jha, A. (eds) (2025). *Artificial Intelligence for Cyber Security and Industry 4.0*. CRC Press, Boca Raton, FL.
- Sokhansanj, B.A. (2025). Local AI governance: Addressing model safety and policy challenges posed by decentralized AI. *AI*, 6(7), 159.
- Srikumar, M., Chang, J., Chmielinski, K. (2024). Risk mitigation strategies for the open foundation model value chain. *Partnership on AI*.
- Widder, D. (2023). Ethics down the AI supply chain: Playing with power. Doctoral Thesis, Carnegie Mellon University.

