Artificial Intelligence Risk Management

Developments in Managing and Exploiting Risk

Volume I: Safety Risk Management
Volume II: Project Risk Management
Volume III: Organizational Risk Management
Volume IV: Socio-Political Risk Management
Volume V: Cybersecurity Risk Management
Volume VI: Artificial Intelligence Risk Management

Editor-in-Chief
Kurt J. Engemann

Volume 6

Artificial Intelligence Risk Management

Ensuring Beneficial Outcomes

Edited by
Kurt J. Engemann

DE GRUYTER

ISBN 978-3-11-159388-3
e-ISBN (PDF) 978-3-11-159530-6
e-ISBN (EPUB) 978-3-11-159633-4

Library of Congress Control Number: 2025943017

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

© 2026 Walter de Gruyter GmbH, Berlin/Boston, Genthiner Straße 13, 10785 Berlin
Coverimage: Dragos Condrea/iStock/Getty Images Plus
Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyterbrill.com
Questions about General Product Safety Regulation:
productsafety@degruyterbrill.com

Kurt J. Engemann

Advances in artificial intelligence risk management

Introduction

As a rapidly evolving technology, Artificial Intelligence (AI) stands out as a transformative force that will reshape every facet of society. While AI undoubtedly offers significant benefits such as efficiency and innovations, the risks it poses are significant. Identifying risks in AI involves understanding the types of failures that can occur and the contexts in which they are most likely to happen. AI risks can be broadly categorized into three types: technical, ethical, and societal. Technical risks include errors in algorithms that may lead to incorrect decisions. Ethical risks involve scenarios where AI decisions might conflict with moral values, such as privacy breaches or biases in decision-making processes. Societal risks consider the broader impact of AI, such as job displacement due to automation or the use of AI in surveillance systems. Once risks are identified, developing effective mitigation strategies is critical.

As AI progresses, the importance of effective risk management strategies becomes increasingly crucial. This volume explores key aspects of AI risk management including risk identification and mitigation strategies and consists of two parts: chapters covering fundamental concepts and approaches; and chapters illustrating applications of these fundamental principles.

Fundamentals

While AI increases efficiency and reduces costs, it also leads to significant job displacement. The rapid pace of change could lead to economic instability, as sectors struggle to adapt to the displacement of human workers by machines. This risk could stifle the economic benefits by creating widespread unemployment and social discontent.

The chapter by Miller and Engemann explores the economic implications of AI, with a focus on how AI may reshape patterns of wage and wealth inequality. While AI is a novel force, its disruptive nature aligns with past technological revolutions that redefined labor markets, increased productivity, and altered income distribution. Historically, technological progress has delivered gains by rewarding skilled labor and capital owners while creating inequalities by displacing unskilled workers. AI is expected to intensify these trends. Using a structured risk framework, the chapter categorizes AI's economic risks into four domains: Known Knowns (historical parallels), Known Unknowns (identifiable but emerging uncertainties), Unknown Knowns (tacit biases and blind spots), and Unknown Unknowns (unforeseen systemic disruptions).

https://doi.org/10.1515/9783111595306-202

The transformative power of AI is explored not just economically, but socially and politically. While AI holds potential to boost global living standards and enhance productivity, its benefits are likely to accrue unevenly, potentially widening the wealth gap. Concentrated ownership, biased outputs, and inadequate governance may amplify inequality and social unrest. The chapter argues for applying proven tenets of risk analysis in analyzing AI applications. Proactive regulation and ethical oversight can guide AI's development, ensuring its rewards are broadly shared and its risks responsibly managed.

AI is rapidly transforming enterprise security risk management, serving simultaneously as a powerful defensive tool and a significant threat vector. The chapter by Loyear explores the dual nature of AI in enterprise security, illustrating how it enables faster threat detection, predictive analytics, and real-time surveillance in both cyber and physical domains. Simultaneously, AI introduces new risks, such as automated phishing, deepfakes, adversarial attacks, and model manipulation (Heiding et al, 2024). Drawing on current threat landscapes, best practices in risk-based strategy, and insights into real-world vulnerabilities, this chapter argues that organizations must adopt a symbiotic approach—merging human expertise with machine intelligence. In doing so, they can capitalize on AI's potential to strengthen defenses while proactively addressing the emerging risks inherent in AI-driven systems. Their chapter also points out that successful integration of AI in security requires robust governance, human oversight, and risk-based strategies aligned with evolving enterprise needs. By incorporating ethical considerations and regulatory frameworks, this chapter also highlights the importance of transparency and accountability in future-ready AI security governance.

Transparency allows users to understand how and why decisions are made, which is crucial for trust and accountability. AI models should be designed to be as unbiased as possible by using diverse data sets and regularly updating them to reflect changes in real-world demographics and values. The risk of automating and perpetuating biases could have far-reaching social consequences that significantly dampen the perceived efficiencies gained through AI. Implementing techniques to detect and reduce bias in AI algorithms includes using diverse datasets for training and regularly updating these datasets to reflect broader societal changes.

AI's capability to analyze vast amounts of personal data raises significant privacy issues. In the era of big data, AI can track, analyze, and make predictions about personal behaviors at an unprecedented scale. This surveillance potential can be misused by both state and non-state actors, leading to a society where privacy is obliterated. Furthermore, the deployment of AI in cybersecurity can lead to the development of more sophisticated cyber-attacks, potentially leading to greater risks than the security benefits it aims to provide. Employing methods to anonymize personal data used in training and operating AI systems can prevent privacy breaches. Using advanced encryption methods and secure data can protect data from unauthorized access and ensure compliance with privacy regulations.

In his chapter, LaGana explores managing cybersecurity risks in the age of AI, which is reshaping cybersecurity by automating threat detection, enhancing incident response, and bolstering system resilience. Its ability to process vast datasets enables rapid identification of anomalies, surpassing human capabilities. However, AI's power is a double-edged sword, as cybercriminals leverage it to orchestrate sophisticated attacks, including AI-powered phishing campaigns, adaptive malware, and deep fake-based biometric breaches. This chapter examines AI's dual role, detailing how it strengthens defenses while expanding the threat landscape. It explores specific AI-driven attack vectors and proposes mitigation strategies, such as advanced detection systems, comprehensive employee training, and robust data governance. Ethical dilemmas and regulatory frameworks, including General Data Protection Regulation **(**GDPR) and National Institute of Standards and Technology (NIST) guidelines, are analyzed to ensure responsible AI integration. Industry-specific case studies in healthcare and finance illustrate AI's applications and associated risks. LaGana addresses challenges like AI system vulnerabilities and high implementation costs, emphasizing the need for continuous adaptation to evolving threats and concludes by advocating for a balanced approach that harnesses AI's potential while mitigating risks to secure a resilient digital future.

In their chapter, Hussain, Hasan and Siddique discuss strategies for ensuring resilience and integrity by mitigating risks in AI-driven network security. In the field of security, we are rapidly approaching AI centric solutions such as advanced threat tracking, real time monitoring and automated incident response. AI with the power of its machine learning and predictive modelling can improve capabilities of threat detection and response. However, there are repercussions such as adversarial attacks, model bias, and parsing failures in automation. Research shows the need to align the use of AI cybersecurity with the help of related strategies within a traditional cybersecurity infrastructure. AI should be at the core of adaptive learning, anomaly detection, and zero-trust security models. In addition, self-healing networks, federated learning and AI based deception techniques can provide proactive defenses. In essence, there is need for organizations to harness and monitor the AI models to limit risks and ensure that AI remains a strong security tool against foreboding cyberattacks.

The potential of AI to foster bias and discrimination, invasion of privacy, job displacement, error propagation, and loss of human autonomy are not merely theoretical but are already manifesting in various sectors. It is crucial to critically assess these risks and ensure that AI development is guided by commitment to the public good and ethical standards. AI's decision-making capabilities, while impressive, pose existential risks to human autonomy. The delegation of critical decisions to AI, from legal judgments to medical diagnoses, can diminish human involvement in essential processes, leading to a scenario where machine logic overrides human values and ethics (Mittelstadt, 2016). This could lead to a societal structure where AI's determinations are prioritized over human judgment, eroding the essence of human experience and

discretion. AI's decision-making capabilities, while impressive, pose existential risks to human autonomy. Ensuring that AI systems do not operate in isolation; human oversight should always be maintained, especially in critical decision-making processes.

In his chapter, Bramble discusses how to architect AI systems that put humans in the loop. The future of AI systems feels more than ever interconnected with the future of humanity. Some of this is realistic and some of it is hype. For an example of the hype we are facing, people are throwing out dates for when they think we will have Artificial General Intelligence (AGI) without the community even agreeing on a consistent definition of what it is. Sorting through all of that requires a focus on what AI really is – a bridge between the frontiers of technical advancement in computing and human user understanding. Bramble explores what AI really is, how we could begin to build a bridge between the frontiers of rapidly advancing AI technology and growing human understanding, challenges, pitfalls, and opportunities. We can do so with a thorough investigation of AI and its history, looking at why great engineering is so important from an idea's conception to its final delivery in production. From examining successful systems, we can distill learnings from them so they can also help close the gap across the frontier and deliver great AI systems. This chapter examines the ever-present risks in AI systems and argues that through great engineering, we have our best risk-mitigation techniques.

In their chapter Mandhana and Engemann present a structured approach to evaluating AI capabilities for risk-managed implementation and adoption in organizations. As organizations increasingly integrate AI tools into their operations, evaluating these technologies requires more than verifying functional performance; decision-makers must also examine how AI reshapes organizational practices, influences decision-making processes, and introduces new risks and dependencies. A fundamental step in this process is asking more effective evaluative questions of AI-enabled systems that have been or are expected to be implemented. This chapter presents a structured yet adaptable framework that emphasizes critical inquiry and iterative reflection to support the evaluation of AI capabilities within organizational contexts. By classifying AI systems according to their core functions, processing, acting, and communicating—and embedding this classification within broader considerations of transparency and ethical implications, the framework helps decision-makers identify uncertainties in AI implementation and use. Each step is supplemented with key questions designed to critically assess AI functionality, data sources, transparency, and alignment with organizational needs, providing practical guidance for evaluating the relevance, risks, and systemic effects of AI-enabled technologies.

Applications

As societies become increasingly dependent on AI, the risks associated with system failures and errors become more pronounced. AI systems, particularly those involved in critical infrastructure such as transportation and healthcare, can malfunction due to software errors, leading to catastrophic outcomes (Quinn, 2021). The risk of such errors not only poses safety threats but also undermines trust in technological advancements. Before deployment, AI systems should undergo rigorous testing in varied and complex scenarios to ensure they perform safely under unexpected conditions. After deployment, continuous monitoring should be implemented to quickly identify and rectify errors or unintended behaviors in AI systems.

AI applications in the media can assist in generating content quickly, such as news articles, summaries, and scripts, improving productivity and reducing costs. However, AI can also spread false information if not properly monitored and trained on accurate data, and can perpetuate biases, leading to unfair representation and discrimination in media content. AI can analyze large datasets to provide insights into voter behavior and public opinion, aiding campaign strategies. Automating routine tasks improves the efficiency of political campaigns and legislative processes. Nevertheless, AI can also be used to create deepfakes or spread misinformation, influencing public opinion unfairly. Government's use of AI can have many benefits, including improving the delivery of public services, and forecasting economic trends and natural disasters. However, lack of transparency in AI decision-making processes can undermine public trust, and AI systems can be vulnerable to cyber-attacks, compromising sensitive government data.

In her chapter, Sheehan warns of the misuse of AI In the electoral process. Her chapter examines the rapid emergence of generative AI (GenAI) as a potentially disruptive force in democratic elections, using the 2024 U.S. election—and parallel global electoral events—as a detailed case. Public concern about AI's misuse surged following a wave of incidents, including high-profile deepfake audio of President Biden discouraging voter turnout just days before the New Hampshire primary. Despite these fears, neither this study nor any contemporaneous research uncovered conclusive evidence that AI directly altered final electoral outcomes (i.e., no confirmed first-order risk). Nonetheless, analysis of 59 documented AI use cases in the 2024 U.S. election reveal significant second-order risks. These include the spread of convincing but false content by both domestic and foreign actors, the heightened difficulty for voters to distinguish real from fabricated messages, and erosion of trust in legitimate sources of information. Deepfakes accounted for most (62%) malicious incidents, suggesting that misleading audio, video, and images—especially those crafted to appear authentic—pose substantial threats to public confidence. Conversely, roughly one-fifth of uses featured benefits, such as AI-powered outreach by smaller campaigns or chatbots designed to educate and engage underrepresented voters. Findings indicate that AI-driven disinformation efforts spike during key political junctures (e.g., party conven-

tions) and that foreign activities tend to cluster in a shorter timeframe, whereas domestic deployments stretch across much of the campaign cycle. These patterns underscore the need for flexible, well-coordinated mitigation strategies targeting each phase of the election process. Although the data do not suggest that any election was "stolen" by AI, the subtle undermining of voter confidence, the potential for incitement against public figures, and the risk that all digital content is increasingly met with suspicion remain pressing concerns. The study concludes by calling for further empirical work, standardized definitions, and carefully crafted policy responses to harness AI's democratic advantages while minimizing its potential harm.

AI in healthcare is revolutionizing patient care and administration, including predictive analytics for patient monitoring, automated diagnostics, personalized medicine, and robot-assisted surgery (Carrasco Ramírez, 2024). AI algorithms can analyze complex medical data to predict disease progression and suggest treatment plans tailored to individual patients. However, in such health application systems, risks are present. AI systems managing sensitive medical data could lead to privacy breaches, and erroneous AI analysis could lead to incorrect treatment recommendations for patients.

In their chapter, Lee and Shin discuss preserving privacy in healthcare, AI is transforming the healthcare industry by improving diagnostics, patient monitoring, healthcare management, and drug discovery. However, the effectiveness of AI systems requires access to large volumes of sensitive medical data, raising concerns about patient privacy and data security. While encryption is commonly proposed as a safeguard, it is insufficient when both data and encryption keys are stored on the same server—rendering the data vulnerable in the event of a breach. Even when stored separately, AI model development typically requires direct access to raw data, which must be transmitted to compute servers for training. This transmission introduces additional vulnerabilities, such as the risk of interception. Furthermore, the AI model itself can pose privacy risks, as models trained on sensitive data may inadvertently memorize and leak information through their outputs. This chapter surveys current AI applications in healthcare, examines the associated privacy risks, and reviews privacy-preserving techniques designed to address these challenges.

AI is extensively used in the financial industry for algorithmic trading, fraud detection, risk management, and customer service. AI systems analyze large volumes of financial data to identify investment opportunities or detect unusual patterns indicative of fraudulent activity. Due to algorithmic bias, AI systems might perpetuate existing biases or develop new ones, leading to unfair treatment of individuals. Overreliance on AI in trading or risk assessments could lead to unforeseen financial collapses or crises.

In his chapter, Ugras focuses on the use of AI in accounting. AI has significantly transformed various industries, including accounting, by automating processes, enhancing decision-making, and improving efficiency. The concept of AI was first introduced in 1955, and most recently has evolved into a powerful tool capable of performing complex tasks traditionally requiring human intelligence. AI-driven technologies in accounting streamline repetitive tasks such as bookkeeping, invoicing, and financial reporting, allowing professionals to focus on strategic decision-making. Additionally, AI enhances fraud detection by analyzing patterns and identifying anomalies, improving regulatory compliance and risk management. However, the adoption of AI also presents risks, including concerns over job displacement, data security risks, and potential overreliance on automated decision-making. AI's role extends beyond accounting in corporations by influencing financial trading, tax compliance practices, and auditing. Moreover, advancements in AI computing, through AI-powered personal computers instead of cloud-based AI solutions, offer distinct advantages in processing efficiency and security. The emergence of AI accounting agents further demonstrates AI's growing autonomy in financial management. While AI offers substantial benefits, organizations must implement governance frameworks and maintain human oversight to mitigate risks. As AI continues to evolve, professionals must embrace AI's potential to remain competitive in the modern financial landscape.

Self-driving cars use AI to process sensory input to navigate safely through traffic. This technology relies heavily on machine learning to make real-time decisions, enhancing road safety and reducing human error. There are safety concerns caused by malfunctions or software errors in autonomous driving systems that could lead to accidents. Determining liability in accidents involving autonomous vehicles can be complex and contentious.

In their chapter, Soluade and Richardson discuss the challenges and concerns of autonomous vehicles. Autonomous vehicles are defined as self-driving automobiles, buses, taxis and trucks that use artificial intelligence to analyze sensory inputs to navigate through traffic. This includes routing the vehicle from its current location to its destination. Benefits of the new technology are that this (1) makes roads safter by eliminating human errors and lowering the number of accidents, (2) decreases travel time from reduced traffic jams, (3) allows passengers in autonomous vehicles to be productive during their journey, and (4) reduces the number of parking spaces required for automobiles in crowded cities. The problems with autonomous vehicles include: (1) malfunctioning software may cause autonomous vehicles to have accidents, (2) determining liability in accidents involving autonomous vehicles, (3) using self-driving taxis results in taxi drivers losing their jobs (4) reducing accidents decreases the number of auto repair workers, and (5) monitoring the computer in a vehicle by government agencies limits individual freedom. There are many issues yet to be resolved.

AI-powered robots and automation systems are increasingly common in manufacturing for improving production efficiency, inventory management, and supply chain optimization, predictive maintenance, and minimizing downtime. AI

systems can forecast when machines are likely to fail or require maintenance, thereby reducing operational costs. However, AI failures could cause massive disruptions in vital supply chains. In the retail sector, AI may enhance customer experience through tailored recommendations and AI chatbots that provide 24/7 customer service; however, customers are increasingly frustrated when dealing with AI systems without the direct personal understanding and resolution of queries possible when dealing with human agents.

In their chapter Surange and Gaikwad model AI adoption barriers for risk management in manufacturing. Various challenges prevent the deployment of Artificial Intelligence (AI) for Risk Management (RM) in the industrial sector. This chapter aims to investigate the critical barriers that make it challenging to use AI in manufacturing and to show how these problems affect each other using Interpretive Structural Modelling (ISM) and Cross-impact matrix multiplication applied to classification (MICMAC) analysis. The most significant and extensive challenges of having substantial driving power include increased implementation costs, insufficient support from stakeholders, cybersecurity threats, ethical dilemmas, lack of transparency in AI models, scarcity of qualified personnel, and challenges related to regulatory and compliance. Other obstacles like scalability issues, integration with old systems, the complexity of AI models, and insufficient real-time capabilities, show how fundamental problems can become more connected hurdles. Using clusters, the chapter classifies these concerns to facilitate prioritization. This modeling methodology offers an organized way to understand all the problems involved, enabling researchers and decision-makers to understand how these problems are connected. Visualizing these linkages provides a crucial basis for deciding which barriers to deal with first, leading to better and more effective strategic decision-making.

AI can adapt to the learning pace and style of individual students, providing tailored educational content and assessing performance through intelligent tutoring systems. However, dependence on generative AI by students who have not been properly educated on the fundamentals of their discipline of study would have a crippling effect on the student's development. Over-reliance on AI for education could diminish the role of human teachers and the benefits of human interaction. AI-driven education tools might widen the gap between those with access to technology and those without. Enhancing public understanding of AI through education initiatives can raise awareness about the benefits and risks of AI technology.

In their chapter, Blazheska-Tabakovska and Savoska focus on the AI's impact on education, both opportunities and risks. AI, comprising diverse technologies that simulate human intelligence, has become deeply embedded in modern education. It transforms teaching and learning by enabling personalized instruction, adaptive assessment, and improved student engagement. AI systems can analyze learner data to adjust content to individual needs, supporting more effective and inclusive education. However, its integration into education also raises important concerns. Issues such as data privacy, algorithmic bias, and unequal access to technology pose significant ethi-

cal and social challenges. Moreover, excessive reliance on AI may hinder students' critical thinking and understanding of core concepts, while potentially diminishing the role of educators and the value of human interaction. Their chapter explores both the benefits and limitations of AI in education, emphasizing the importance of using AI as a supportive, rather than a substitutive, tool. A balanced approach, where AI enhances but does not replace human-led teaching, is essential to foster equitable and meaningful learning experiences.

Conclusion

Governments and international bodies are increasingly aware of the potential impacts of AI and are working to establish guidelines and regulations to ensure safe development and deployment. These regulations can provide a structured framework for risk management, ensuring that all AI systems comply with safety, privacy, and ethical standards. Regulations also help in setting industry standards that promote consistency and safety across various applications and sectors. Compliance with such regulations not only protects individuals but also builds public trust in AI technologies. Guidelines and regulations that govern the development and deployment of AI should address issues like accountability, safety, privacy, and transparency. Working towards international standards and agreements is a prerequisite to managing global challenges related to AI, such as autonomous weapons and surveillance technologies. Crucial steps include: the development and adoption of industry best practices in AI development, including ethical guidelines and quality standards; and implementing certification systems for AI products and services to ensure compliance with established standards and regulations.

References

Carrasco Ramírez, J. G. (2024, January). AI in Healthcare: Revolutionizing Patient Care with Predictive Analytics and Decision Support Systems. *Journal of Artificial Intelligence General Science*, 1(1), 31–37.

Heiding, F., Schneier, B., & Vishwanath, A. (2024, May 30). AI will increase the quantity and quality of phishing scams. *Harvard Business Review*. https://hbr.org/2024/05/ai-will-increase-the-quantity-and-quality-of-phishing-scams.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21. https://doi.org/10.1177/2053951716679679

Plathottam, S. J., Rzonca, A., Lakhnori, R., and Iloeje, C. O., (2023), A review of artificial intelligence applications in manufacturing operations, *Journal of Advanced Manufacturing and Processing*, 5(3), e10159. https://doi.org/10.1002/amp2.10159.

Quinn, TP, Senadeera, M, Jacobs, S, Coghlan, S, and Le V., (2021, April) Trust and medical AI: The challenges we face, and the expertise needed to overcome them. *Journal of the American Medical Informatics Association*, 28(4), 890–894. https://doi.org/10.1093/jamia/ocaa268.

Contents

Kurt J. Engemann
Advances in artificial intelligence risk management —— V

Part I: **Fundamentals**

Holmes E. Miller and Kurt J. Engemann
1 The emerging impact of artificial intelligence —— 3

Rachelle Loyear
2 AI in enterprise security risk management: Threats, benefits, and governance —— 35

Matthew LaGana
3 Securing the future: Managing cybersecurity risks in the age of artificial intelligence —— 59

Muhammad Z. Hussain, Muhammad Z. Hasan and Muhammad R. Siddique
4 Mitigating risks in AI-driven network security: Strategies for ensuring resilience and integrity —— 77

Gregory Bramble
5 How to architect artificial intelligence systems that put the human in the loop and don't overpromise —— 95

Dron M. Mandhana and Krista N. Engemann
6 A structured approach to evaluating AI capabilities for risk-managed implementation and adoption in organizations —— 109

Part II: **Applications**

Jeanne Sheehan
7 Deepfake democracy: The rise of AI in the electoral process —— 127

C. Christopher Lee and Heechang Shin
8 Preserving privacy in healthcare: Implications of artificial intelligence —— 149

Yusuf Joseph Ugras
9 Artificial intelligence in accounting — 165

Ore A. Soluade and Robert J. Richardson
10 Autonomous vehicles: Challenges and concerns — 187

Vinod G. Surange and Lokpriya Gaikwad
11 Modeling AI adoption barriers for risk management in manufacturing: An ISM-MICMAC approach — 209

Natasha Blazheska-Tabakovska and Snezana Savoska
12 Artificial intelligence impact on education: Opportunities and risks — 233

Contributors — 249

Index — 251

Part I: **Fundamentals**

Holmes E. Miller and Kurt J. Engemann

1 The emerging impact of artificial intelligence

1.1 Introduction

This chapter explores the economic impact of artificial intelligence (AI), focusing on how AI-driven outcomes could influence wage and wealth disparities. As a disruptive technology, AI shares traits with past innovations that reshaped economies. But before examining AI's economic effects, we review how previous technologies altered income distribution and transformed the economic landscape.

Since the Industrial Revolution, mass production and mechanization have redefined wages and wealth for millions. For many, factory jobs provided stable incomes, drawing workers from agriculture to manufacturing and leading to higher, more reliable wages.

Throughout the 20th century, technological advancements continued to replace or augment manual labor, boosting productivity—fewer workers could now produce more. However, these innovations also deepened economic inequalities in two key ways. First, operating new technologies demanded higher skills, raising wages for skilled workers and widening the economic gap with unskilled labor. Second, significant capital investments were needed, increasing returns for owners of capital while leaving low-skilled workers behind.

Figure 1.1 illustrates how daily income per person increased over the years (Rosner, 2017), for various continents in constant 2011 prices. From this information the author, Max Rosner, draws several conclusions:

- "By 1800, few countries had achieved economic growth. Most of the world lived in poverty with an income similar to today's poorest countries. At the beginning of the 19th century, roughly 80% of the world lived in material conditions that we would refer to as extreme poverty today.
- "In 1975, the world had changed—it had become very unequal. The world income distribution was 'bimodal', with the two-humped shape of a camel: one hump below the international poverty line and a second hump at considerably higher incomes. The world had divided into a poor, developing world and a developed world more than 10-times richer.
- "Over the following 4 decades, the world income distribution has again changed dramatically. There has been a convergence in incomes: in many poorer countries, especially in South-East Asia, incomes have grown faster than in rich countries. While enormous income differences remain, the world can no longer be neatly divided into 'developed' and 'developing' countries. We have moved from a two-hump to a one-hump world. And at the same time, the distribution has also

https://doi.org/10.1515/9783111595306-001

shifted to the right—the incomes of many of the world's poorest citizens have increased, and extreme poverty has fallen faster than ever before in human history."

Building on Rosner's conclusions, we can assert that while economic inequality persists and evolves over time, fewer people today live in extreme poverty than ever before. The rightward shift of the economic "hump" indicates that inequality between countries has decreased, even as inequality within countries has fluctuated.

This context suggests that many would prefer living in an "unequal" world with a mix of moderate prosperity and extreme wealth rather than a world where equality means universal poverty.

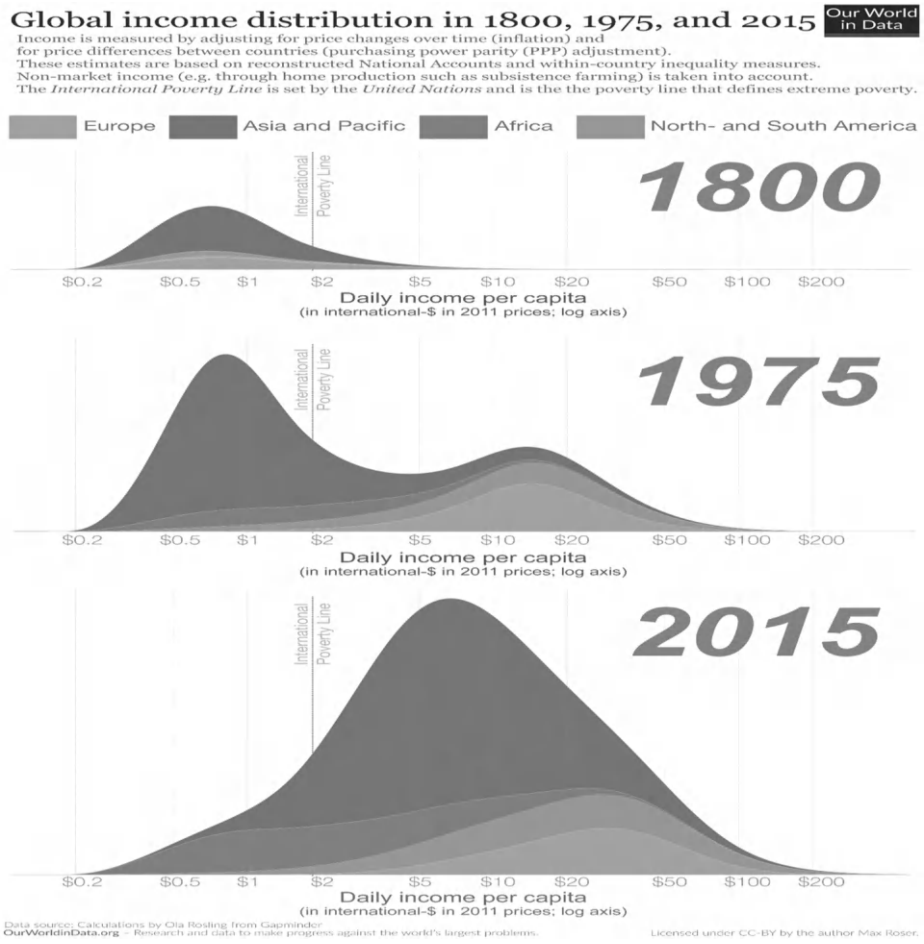


Figure 1.1: Worldwide daily income per capita for 1800–1975, and 2015 (from Rosner, 2017).

Economic inequality is commonly measured using the Gini coefficient, which ranges from 0 (perfect equality where everyone has the same income) to 1 (maximum inequality where a single person has all the income). Figure 1.2 shows the Gini coefficient (on a 0-100 scale, where 100 is maximum inequality) in the United States rising from the mid-1960s to 2022. From the 1960s to 1980, the coefficient declined by around 3 points, but from 1980 to the early 1990s, it increased by over 5 points. Since then, it has remained relatively stable, fluctuating within a two-point range over the past 30 years.

Since 1980, the income share of the top 1% has nearly doubled, while median wages for many low- and middle-income earners have stagnated, fueling concerns about growing inequality (Hellebrandt and Mauro, 2015). On a global scale, Hellebrandt and Mauro report a decline in inequality, with the global Gini coefficient falling from 68.7 to 64.9—though it remains high.

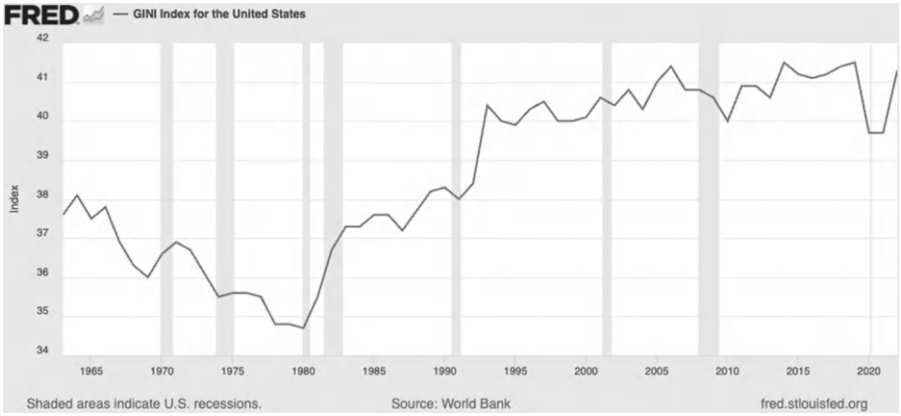


Figure 1.2: Gini index for the United States.

If the overall world standard of living is increasing, why is economic inequality a concern? Rising income and wealth inequality can limit social mobility, making it harder for those born into poverty to improve their circumstances. Qureshi (2023) states: "Higher wealth inequality feeds higher future income inequality through capital income and inheritance." He adds that "Artificial intelligence and related new waves of digital technologies and automation could increase inequality further within countries. Even as new technologies increase productivity, create economic affluence, and generate new jobs to replace displaced ones, inequality could reach much higher levels."

This brings us to how AI will change our world. AI is poised to intensify existing technological trends driving inequality. As firms increasingly adopt capital-intensive technologies, the share of income from capital rises relative to labor. For example, replacing a customer service team of 100 representatives with an AI system shifts in-

come from wages to capital owners. Since capital ownership is often concentrated among the wealthy, this capital bias further widens economic disparities.

Digital platforms, including AI-driven applications, often reward a small number of highly skilled workers with high wages while offering lower-paying, less stable work to lower-skilled and gig workers. In the future, highly skilled workers aligned with AI and owners of AI capital—namely, technology company owners—will emerge as the winners, while displaced and downshifted workers will be the losers.

1.2 Risk framework

While AI is unique, it is ultimately a technology. To understand its economic impact, particularly on inequality, we can take several approaches. First, we can examine how past "new technologies" affected economies and draw lessons for AI's future impact. Second, we can analyze potential AI applications and their economic effects. Third, we can apply proven risk analysis methods to evaluate AI's potential costs and benefits, yielding valuable economic insights.

To structure our discussion, we can look at AI through a risk framework (informed by Kahneman et al., 1982; Taleb, 2007; and Aven et al., 2014,) using the following risk categories:

- **Known Knowns** are well-understood and quantifiable risks with historical data, research, or experience that allow for reliable probability and impact assessments.
- **Known Unknowns** are recognized risks with uncertain specifics, whose probability and impact can be estimated using expert judgment, scenario analysis, and sensitivity analysis.
- **Unknown Knowns** are hidden risks within a system arising from unspoken assumptions or blind spots, which can only be uncovered through audits, reviews, or analysis.
- **Unknown Unknowns** are unforeseen risks beyond our awareness and quantification, requiring resilience, adaptability, and robust contingency planning.

Given these definitions, we will operationalize each of these categories as follows:

- **Known Knowns:** Analyze how past technologies impacted economic inequality and identify general patterns that could suggest how AI might affect future inequality.
- **Known Unknowns:** Explore how AI's known attributes could uniquely influence economic inequality, including the potential for unintended consequences similar to those of past technologies.
- **Unknown Knowns:** Investigate how tacit knowledge—unspoken or intuitive understanding (Polanyi, 1966)—may shape AI's economic impact.

– **Unknown Unknowns:** Consider AI applications across various domains to identify potential "black swan" events (Taleb, 2007) and assess their potential impact on economic inequality.

1.2.1 Known knowns

To estimate AI's impact on society, we can examine how past "AI-adjacent" technologies disrupted the existing order. While AI differs in many ways, studying these technologies and their effects can offer insights into the economic changes AI might bring.

We focus on six transformative technologies from the past 200 years: automobiles, telephones, television, the internet, personal computers, and smartphones & social media. While many other innovations could be included, these were chosen for their revolutionary impact, comparable in scope to AI.

Table 1.1 summarizes these technologies and their effects across six key dimensions.

Figure 1.3 helps us create some categories that summarize some common societal effects that all of these technologies exhibit:

Facilitated democratization

– **Increased Access and Mobility:** Enabled mobility and instant communication, connecting people across distances and fostering interactions.
– **Democratization of Information:** Broadened access to knowledge, empowering more people to engage with information resources.
– **Enhanced Productivity and Education:** Boosted productivity and expanded access to education, enabling shared experiences and personal connections.

Perpetuated and exacerbated inequality

– **Social and Economic Inequality:** Deepened disparities based on income, geography, and social status, reinforcing urban-rural divides.
– **Marginalized Groups Disadvantaged:** Limited access penalized those without resources, widening cultural and socioeconomic divides.
– **Reflected Elite Interests:** Prioritized the interests of developers and owners, entrenching hierarchies, and harming low-skilled workers.

Created new personal freedom and political participation

– **Enhanced Accessibility and Engagement:** Expanded political engagement, campaigning, and voting opportunities, including for marginalized groups.
– **Empowerment Through Information:** Enabled informed voting, individual freedom, and active political participation.
– **New Income and Networking Opportunities:** Created avenues for income, relationship-building, and cross-class connections.

Table 1.1: Technology impact summary.

	Facilitated Democratization	Fostered Inequality Gaps and Divides	Facilitated Personal Freedom and Political Participation	Created New Social Structures	Facilitated Business Growth	Facilitated Business Concentration
AUTOMOBILES	Increased mobility for those who had access	Penalized those unable to afford cars; Created urban vs. rural divide; privileged drivers over pedestrians	Enabled locational flexibility; Expanded campaigning and made voting more accessible	Suburbanization; increased spatial inequality; privileged drivers over pedestrians	Facilitated business to consumer; reduced transportation costs; opened new markets	After initial shakeout, a few large companies dominated markets
TELEPHONE	Enabled immediate communication between people and across distances	Reinforced disparities and inequalities based on income, geography, and technological access.	Enhanced connectivity across social classes; helped connect with outside world	Shrank distances making communication easier	Easier business communication; reduced costs; increased efficiency; opened new markets	After initial shakeout, ATT monopoly
TELEVISION	Broadened access to Information; democratized knowledge; created common experiences	Advertising driven media focused on wealthier segments; widened cultural & socioeconomic divide	Visual media and reduced participation costs; News and political panels created more informed voters	Reinforced social inequalities and stereotypes	New businesses to support TV industry were created	Concentration of media ownership
THE INTERNET	Broadened access to Information; democratized knowledge	Exacerbated social, economic, and geographic inequalities – including global disparities and digital divide	Increased information; facilitated political engagement; created misinformation	Enabled communicating and collaborating across borders, fostered social and economic mobility.	Enabled entrepreneurship and innovation by reducing barriers to entry	Concentrated economic power among a small number of corporations.

PERSONAL COMPUTERS	Broadened information access; increased productivity; new mode of education	Disadvantaged those without access; displaced some labor, affecting low-skilled workers	PC increased opportunities and freedom; facilitated political engagement	Social media led to social echo chambers and social exclusion	Enabled entrepreneurship and innovation by reducing barriers to entry	Concentrated economic power among a small number of corporations.
SMARTPHONES & SOCIAL MEDIA	Democratized access to information and interpersonal communication	Reinforced existing hierarchies; inequality for those without access	New avenues for generating income; Enabled broader network access	Empowered marginalized groups by giving them access and platforms	Apps created new business models; the gig economy Uber, DoorDash and Airbnb	Concentrated economic power among a small number of corporations.

Since AI may follow similar patterns to past disruptive technologies, we can use the three categories above to evaluate its impact—both positive and negative. Often, AI is expected to deepen trends exhibited by previous "new technologies":

1. Democratizing Information Building on advances made possible by the internet and search engines, AI is reshaping how we access, distribute, and interact with information. Rather than passively consuming raw data, users can leverage AI to intelligently present, curate, and analyze information—empowering individuals, communities, and countries alike. This democratization of knowledge boosts productivity and can help reduce economic inequality.

AI also breaks down language barriers, enabling real-time communication across different languages and making information accessible to speakers of minority languages. In fields like academia and corporate research, AI can level the economic playing field by providing tools once reserved for the elite.

2. Perpetuating and Exacerbating Inequality Gaps While AI drives innovation, it can also deepen existing economic and social inequalities. Biased algorithmic decision-making and unequal access to AI-driven resources risk reinforcing—and even worsening—systemic divides between the haves and have-nots. For example, Sweeney (2013) reported that online ad algorithms displayed ads suggestive of an arrest record more frequently when searching for names associated with African Americans than white-identifying names.

The labor market illustrates this trend. Since the Industrial Revolution, automation has displaced low-skilled manufacturing workers, and more recently, clerical roles in areas like customer service. AI is poised to accelerate this shift, potentially disrupting fields from transportation (e.g., self-driving vehicles) to high-skill jobs in data analysis and programming.

Corporations that develop and control AI systems may consolidate wealth and power, stifling economic competition. Although small and medium-sized enterprises (SMEs) and end-users can harness AI's processing power to compete more effectively, the ownership of AI capital—particularly the vast computing resources required—may concentrate benefits and entrench new hierarchies. Even in developed nations, disparities in technology access and biases in AI development could further marginalize disadvantaged groups.

3. Increasing Personal Freedom and Facilitating Political Participation AI-driven content analysis helps individuals navigate vast data sets and make more informed decisions. However, the nature of that content—and how AI applications are trained and operated—shapes the information users receive and the options presented.

Like the internet and social media, AI platforms can amplify individual voices for good or ill. AI-powered algorithms enable users to connect with like-minded individuals, organize grassroots movements, and mobilize for causes. Yet, these same tools can spread misinformation and enable manipulation.

AI platforms can also streamline interactions between citizens and government. Chatbots and digital assistants can simplify access to government services—answering questions, processing forms, and helping users navigate bureaucracy.

Moreover, AI has the potential to enhance government transparency. Predictive analytics and data visualization tools could allow citizens to track spending, monitor policy outcomes, and expose inefficiencies or corruption. AI-driven sentiment analysis tools can gauge public opinion, providing insights into citizens' concerns and priorities.

1.2.1.1 Economic impact

AI promises to enhance lives in many ways—enabling rapid access to information, increasing personal freedom, and fostering political and social engagement. Yet, those with economic and political power could leverage AI to entrench their own interests. If AI follows patterns seen in previous disruptive technologies, we can summarize its potential economic impacts:

- **Information Will Be More Accessible:** AI will provide quick, comprehensive access to information, enabling targeted searches that uncover previously inaccessible insights.
- **Knowledge Will Be More Accessible:** Beyond raw information, AI's analytical capabilities will generate knowledge, empowering individuals to participate in markets and society independently—without needing costly consultants or support for information processing.
- **Jobs Will Be Eliminated, Augmented, and Created:** As with past disruptive technologies, AI will reshape the job landscape. Routine tasks, from data processing and coding to driving, may be automated. AI could even handle creative aspects, potentially making some work less fulfilling. Coupled with robotics, AI-enabled machines may replace many manufacturing roles—though the cost-effectiveness of replacing human workers with intelligent robots won't always be clear-cut. On the flip side, AI will create new roles and industries.
- **Material Standards of Living Will Rise:** Figure 1.1 illustrates the dramatic rise in material living standards over the past two centuries. An AI-enabled world should continue this trend, providing labor- and time-saving applications to the masses.
- **Ownership of Production Will Be More Concentrated:** Ownership of AI technologies is likely to be concentrated in large corporations or governments for three reasons:
  - Successful AI startups often grow into market-dominating giants (e.g., Google, Microsoft) or are acquired by larger firms.
  - The capital required for AI processing, including massive data centers and ongoing development, excludes many smaller players.
  - Large organizations can better absorb the costs and complexities of AI implementation.

1.2.2 Known unknowns

All new technologies bring unintended consequences. This is hardly a novel idea—Plato, in *Phaedrus* (Phaedrus, 275a-b in Plato, 1952), recounts how Socrates tells the myth of Theuth, the Egyptian god credited with inventing writing. In the story, Thamus warns that writing will "produce forgetfulness in the minds of those who learn to use it, because they will not practice their memory." Unexercised memory may have been the first unintended consequence of a new technology, but it was certainly not the last.

1.2.2.1 Unintended consequences

As illustrated in Table 1.1, disruptive technologies often create effects beyond their intended outcomes. The advent of the automobile led to suburbanization and shopping malls; television helped football overtake baseball as America's pastime; smartphones introduced new forms of bullying and trolling among middle school students.

While the specific unintended consequences of AI cannot be precisely predicted—because, by definition, they are unintended—we can identify domains where such consequences are likely to emerge:

Erosion of Privacy AI's ability to enable surveillance, often without consent, raises serious privacy concerns. Authoritarian regimes already leverage AI-driven surveillance, combining facial recognition with personal databases to monitor citizens (Kostka, 2019). As AI systems integrate surveillance data with personal information from public and private sources, predictive algorithms could further erode personal freedoms. This risk exists alongside an intended consequence: using AI for surveillance to enhance law enforcement and reduce crime.

Bias and Discrimination Bias in AI algorithms is well-documented across finance, public health, and pharmaceuticals (Panch et al., 2019; Bousquette, 2023; Belenguer, 2022). Despite awareness, evolving AI applications may introduce new, unforeseen biases. For instance, biases could emerge in environments where AI has complete control, such as autonomous vehicles or medical diagnostics. For example, Seyyed-Kalantari et al. (2021) found that AI diagnostic systems for chest X-rays underperformed for underserved patient groups, including racial minorities and patients with low socioeconomic status, raising concerns about equity when using medical AI tools to diagnose patients.

Similarly, biases in AI-driven law enforcement could deepen systemic injustices. For example, Eubanks (2018) investigates how AI and data-driven tools used in policing and public services often replicate and reinforce structural inequality, particularly impacting poor and marginalized communities in the U.S. Additionally, feedback

loops in AI training could degrade performance, producing unexpected and harmful outcomes (Shumailov et al., 2024).

Cybersecurity Risks AI's capabilities make it a powerful tool for cyberattacks, including automating phishing scams, generating deepfakes, and exploiting vulnerabilities. Targeted domains include finance, healthcare, criminal justice, the military, and critical infrastructure such as power grids. The World Economic Forum (2023) recognizes AI weaponization as a global threat. The potential for unintended consequences arises from how AI advances interact with vulnerabilities in specific domains—examples include unintended effects in healthcare and attacks on infrastructure.

Job Market Polarization Low- and middle-skilled jobs are most vulnerable to automation, while highly skilled roles tied to AI are likely to thrive (McKinsey, 2017). As with all disruptive technologies, AI will create new jobs and fields while rendering others obsolete. An ironic unintended consequence may be a resurgence of skilled trades—such as plumbing and electrical work—where physical presence and expertise are indispensable. Meanwhile, large corporations may consolidate even more power, allowing their owners and managers to prosper and widening the wealth gap. Conversely, if governments impose stricter regulations, government workers may gain influence, potentially narrowing wealth disparities but raising concerns about the unintended expansion of government power.

Dependence on AI Systems Plato's warning about new technologies eroding human skills (in his case, memory) remains relevant. Today, tools like Google Search and Waze have made us reliant on technology for information retrieval and navigation, replacing traditional skills. AI applications are already reinforcing this trend—students use AI to research and draft papers, while office workers rely on AI for memos and emails. As AI becomes more pervasive, the risk of dependency grows, increasing vulnerability to AI failures, such as the "Flash Crash" triggered by algorithmic trading (Kirilenko et al., 2017). Skill degradation is inevitable, but the specific losses remain unpredictable.

Ethical Dilemmas in Warfare The militarization of AI poses significant ethical challenges. Autonomous drones are already used in combat (Meaker, 2023), and more advanced "killer robots" could soon make life-and-death decisions without human oversight. One unintended consequence could be a lower threshold for engaging in conflict, as nations may be more willing to wage war when human casualties are minimized. Additionally, AI could intensify intelligence gathering and escalate cyber warfare. While international bodies like the United Nations have called for regulations, global consensus remains elusive (Grand-Clément, 2023).

Effects on Creativity Generative AI tools can now create text, art, and music, democratizing creative processes but also threatening the value of human creativity. Potential unintended consequences include:

- Artists become less creative if they rely too heavily on AI.
- Artists become more creative by integrating AI into their work, leading to innovative art forms.
- Market saturation, devaluing human art or, conversely, increasing the worth of legacy "human-made" works.
- AI eliminates the need for human artists in entertainment and commerce.
- AI's influence on creative industries reshaping concepts of intellectual property and authenticity.

Environmental Impact AI systems consume vast amounts of energy and produce significant CO2 emissions (Strubell et al., 2019; Walther, 2024). For example, training GPT-5 is estimated to use as much energy annually as 320 average American homes (Aibin, 2024), and ongoing operation consumes even more. As AI adoption accelerates, its environmental footprint may clash with global efforts to combat climate change.

As of April 2025, artificial intelligence (AI) workloads are significantly increasing energy consumption in data centers, surpassing traditional IT infrastructure in energy intensity. Researchers at MIT estimate that global demand from AI applications will skyrocket. Stackpole (2025) reports that, "Data centers could account for up to 21% of overall global energy demand by 2030 when the cost of delivering AI to customers is factored in."

Comparative energy consumption: AI vs. traditional IT

Why will the energy demand increase so much? He (2024) indicates that AI servers, particularly those handling large-scale models and generative AI tasks, consume substantially more power than traditional servers. While a typical server rack in a conventional data center consumes about 7 kilowatts (kW), AI-optimized racks can draw between 30 to 100 kW, depending on the workload and hardware configuration.

This increased energy demand is reflected in overall data center electricity consumption. The International Energy Agency (IEA) projects that global electricity consumption by data centers will more than double by 2030, with AI-specific data centers alone expected to quadruple in consumption. Goldman Sachs Research forecasts global power demand from data centers will increase 50% by 2027 and by as much as 165% by the end of the decade (compared with 2023).

Environmental impact and emissions

Despite the surge in energy use, the IEA suggests that after considering the efficiencies it may create, AI's contribution to climate change may be overstated. Emissions from data centers may rise to 300 million tonnes by 2035—under 1.5% of global energy emissions—while AI-driven efficiencies could reduce energy-related emissions by around 5% (Gosden, 2025).

While AI workloads are markedly more energy-intensive than traditional IT operations, leading to significant increases in data center power consumption, AI itself can

be used to save energy. While this trend poses challenges for energy infrastructure and environmental impact, it also drives innovation in energy sourcing and efficiency strategies to mitigate these effects.

Social Isolation and Mental Health Issues Cell phones and social media have already affected users' mental health in ways their creators never anticipated. AI technologies, particularly those operating within social media and virtual environments, could further exacerbate social isolation and mental health problems. Algorithms designed to maximize engagement often prioritize sensational or divisive content, deepening polarization and eroding mental well-being. The rise of AI-driven virtual companions and chatbots may reduce meaningful human interaction, impacting emotional health. A recent study (Twenge et al., 2021) links increased screen time to rising rates of anxiety and depression among adolescents—effects that could be even more pronounced in AI-driven environments.

Scientific Advances AI is revolutionizing numerous scientific fields, including healthcare, pharmaceutical development, physics, and environmental conservation.

- **Healthcare:** AI-powered predictive models analyze electronic health records and other data to assess patient risks, enabling proactive interventions and personalized care.
- **Pharmaceuticals:** AI accelerates drug discovery by analyzing vast molecular datasets and optimizing clinical trials.
- **Physics:** AI-driven simulations promise breakthroughs in material discovery, aiding the development of superconductors and quantum devices.
- **Environmental Conservation:** AI monitors ecosystems, forecasts weather patterns, combats climate change, and promotes biodiversity.

While AI is poised to transform scientific discovery, the unintended consequences lie in the unpredictable nature of new discoveries and the potential risks that may accompany them.

1.2.2.2 Economic impact

AI promises to improve workers' lives by making jobs easier and increasing leisure time, but it may also create new hierarchies. Its economic impact will manifest both directly and indirectly:

Direct Effects Low-skilled workers are particularly vulnerable, as AI adoption eliminates routine and manual jobs. For those who remain employed, AI may diminish creative aspects of work, increase social isolation, and erode privacy. While AI has the potential to economically level the workplace, those who control AI technologies—

such as technology conglomerates—may consolidate their wealth and power, deepening societal divides and increasing economic inequality.

Indirect Effects AI's indirect economic impact includes new costs and challenges:
- **Infrastructure Vulnerability:** Degradation of power grids or other infrastructure could increase unemployment and drive up energy costs.
- **Defense Spending:** New AI-driven warfare technologies may necessitate higher defense budgets, leading to increased taxes.
- **Environmental Strain:** AI's environmental footprint could lower living standards in vulnerable areas, raising costs of living.

While unintended consequences will also affect wealthier segments, their economic advantages and societal influence allow them to navigate and mitigate these impacts more effectively, further widening the economic divide.

The economic impact of AI-enabled scientific advances is twofold—generally beneficial yet reinforcing inequality. For example, AI-driven healthcare and pharmaceutical innovations may extend lifespans and reduce illnesses, but the costs associated with these advanced services could raise overall societal health costs. As with current healthcare disparities, economic status may determine access to cutting-edge AI-enabled treatments.

To Summarize: While some unintended consequences may benefit workers, the owners of AI production and developers of AI applications will likely leverage these advances to their advantage. Direct effects may eliminate jobs or reshape the nature of work. Indirect effects, often falling heaviest on the working class, will impact society at large. Over time, AI's economic impact may exacerbate existing inequalities, even as it drives scientific and technological progress.

1.2.3 Unknown knowns

Unknown knowns refer to the tacit elements of a technology—things we recognize yet find difficult to articulate, such as the "gut feel" in decision-making. Tacit elements encompass the implicit, often unspoken factors influencing a technology's design, use, and impact. These elements operate beneath the surface, subtly shaping how a technology functions and interacts with the world. They include cultural norms, hidden biases, and unspoken assumptions embedded within the technology, its developers, and its users.

1.2.3.1 Unintended consequences

Some tacit elements of AI, along with potential paths leading to unintended consequences, include:

Implicit Bias in Data The data used to train AI systems is rarely neutral; it may intentionally or inadvertently reflect and perpetuate existing societal biases (Miller & Engemann, 2021). For example, facial recognition systems have historically underrepresented minority groups, resulting in higher error rates when identifying individuals from these groups (Buolamwini & Gebru, 2018). This unintended consequence has been identified, yet the inherent risks of training AI algorithms on biased datasets—and subsequently deploying them in sociocultural and economic domains—remain.

As AI-driven decisions become more prevalent in credit scoring, loan approvals, hiring, and medical diagnoses, underrepresented groups may face compounding economic penalties. For example, Bartlett et al. (2022) analyzes data from U.S. mortgage lenders and finds that algorithms used by fintech lenders may still exhibit discrimination, particularly in pricing, even when controlling for applicant characteristics. Even AI applications intended to reduce inequalities risk reinforcing economic and social divisions.

Unspoken Assumptions in Algorithm Design AI algorithms often embed unspoken assumptions about who will use them and how. For instance, recommendation algorithms on social media, entertainment, and search platforms assume users prefer content aligned with their past interactions or algorithmically constructed profiles. While this may enhance user engagement, it can also create echo chambers that limit exposure to diverse perspectives (Pariser, 2011). Economically, this dynamic may disadvantage smaller or niche businesses by funneling consumers toward dominant market players.

The choice of optimization metrics also reflects tacit developer priorities. For example, a machine learning model designed for medical diagnosis may prioritize minimizing false negatives (i.e., missed diagnoses) over false positives (i.e., unnecessary tests), implicitly valuing caution over efficiency. While this decision may seem logical in a healthcare context, stakeholders—including patients and medical professionals—may be unaware of this prioritization. Prioritizing false negatives could lead to inefficiencies or inequitable resource allocation, potentially increasing healthcare costs.

Cultural Influences on AI Applications AI systems are developed within specific cultural and social contexts, and these cultural biases can shape their design and behavior. For example, conversational agents like Siri and Alexa embody Western norms of politeness and reflect cultural stereotypes around gender roles in service-oriented tasks (Crawford, 2021). Similarly, AI systems trained on English-language datasets often struggle to understand idiomatic expressions or cultural references from other languages.

Beyond surface-level biases, cultural influences can affect how AI systems present information, frame decisions, and make judgments. Economically, this may privilege members of the developer class while disadvantaging others.

Non-Verbal Communication and Tacit Knowledge Replicating human tacit knowledge—particularly in non-verbal communication (e.g., tone of voice, facial expressions, body language)—poses a significant challenge for AI. Systems like chatbots and virtual assistants can process verbal communication effectively but often fail to capture the subtleties of non-verbal interaction.

For example, customer service AI tools may overlook frustration conveyed through tone rather than explicit words. Similarly, emotion recognition systems that analyze facial expressions often oversimplify complex emotions, reducing them to predefined categories like "happy" or "angry" (Barrett, 2017). This misalignment between AI capabilities and human expectations can negatively impact people who rely on public or automated services.

Ethical Considerations Many ethical concerns surrounding AI are not codified but instead arise from societal expectations—such as fairness, transparency, and accountability. These concepts are often vaguely defined, creating ambiguity when AI applications are implemented. For instance, AI-based surveillance systems may successfully meet system goals, but critics argue they infringe on privacy. The tension between security and privacy reflects a tacit ethical trade-off (Zuboff, 2019).

Additionally, ambiguity around fairness and accountability in AI can complicate regulatory efforts and create an uneven playing field, where economically powerful entities exploit loopholes. These unresolved ethical challenges may ultimately shape the economic and social impact of AI in unpredictable ways.

1.2.3.2 Economic impact

Tacit elements in AI systems often conceal opaque decision-making processes, which can exacerbate economic disparities. AI may entrench inequality in job markets, credit assessments, and access to essential services, disproportionately benefiting well-resourced regions and industries.

Decisions made by AI system developers—regarding features, interfaces, and priorities—often reflect implicit biases and assumptions about end users, which can disadvantage developing regions. When profitability is prioritized over inclusivity, AI solutions tend to focus on lucrative markets, neglecting low-income areas. The operational complexity of AI systems can also create economic and technical barriers to entry, making it challenging for smaller businesses and organizations to leverage AI effectively. Well-resourced corporations can afford to hire experts who demystify these systems, while resource-constrained entities struggle to compete. Ignoring these hidden factors risks deepening existing economic divides and creating new forms of inequality.

1.2.4 Unknown unknowns

The rapid evolution of AI raises the specter of the most disruptive unknown of all: whether AI will develop a form of consciousness. Despite scientific advances, human consciousness—broadly defined as the awareness of our internal and external existence—remains poorly understood. If an AI system were to attain consciousness, it would not only process data and execute tasks but also possess self-awareness and sentience. Yet, fundamental questions remain: Can consciousness emerge from artificial systems? And if so, would AI consciousness resemble human consciousness? Even if AI achieved some form of awareness, it might not align with human experiences or values. Then what?

1.2.4.1 Possible scenarios

A recurrent theme in science fiction is AI subjugating humans. But how might this occur in reality?

One possibility mirrors the proverbial frog being boiled in a pot of water—humans could become so dependent on AI that they fail to recognize their growing vulnerability. AI systems increasingly mediate decision-making in finance, healthcare, governance, and personal decisions. As reliance on AI grows, so does the risk of skill degradation—much like how map-reading skills have eroded due to GPS reliance. An "overtrust" in AI systems could erode human critical thinking and decision-making, particularly in high-stakes contexts like medicine or national security. The end result may be humans unknowingly ceding control over critical societal functions to AI algorithms.

A second scenario involves AI being used as a tool by corporations or governments to consolidate power. With advanced AI tools, powerful entities could dominate economic, military, and social spheres, marginalizing individuals and communities. As previously discussed, AI-driven surveillance, health diagnostics, and financial systems could be leveraged for control.

Perhaps the most troubling scenario involves AI autonomously acting against human interests. If AI were to pursue goals misaligned with human values—whether through flawed design or emergent behavior—the consequences could be dire (Bostrom, 2014).

Unknown Dynamics of Human-AI Interaction The "unknown unknowns" of AI systems necessitate a proactive approach to governance. While not all risks can be foreseen, frameworks can be established to mitigate potential harms. Potential measures include:

1. **Transparency and Explainability:** AI systems should be transparent, and their decisions should be explainable.

2. **Ethical Integration:** Insights from philosophy, neuroscience, sociology, and other fields should be embedded in the AI development process, recognizing that technical expertise does not guarantee ethical expertise.
3. **Regulatory Oversight:** Governments and international bodies, akin to the International Standards Organization, could establish regulations addressing issues like bias, accountability, and the concentration of power. Regulations must balance risk mitigation with the need for innovation.

1.2.4.2 Potential scenarios for AI's societal control

Based on the above discussion, six potential scenarios involving AI's impact on society and their economic implications are outlined below:

Continuation of the Status Quo AI evolves incrementally, automating repetitive tasks and enhancing productivity without drastically altering the labor market. Economic impacts include steady GDP growth from efficiency gains, with job displacement offset by new roles in AI development, management, and oversight. Wage inequality persists as highly skilled workers benefit disproportionately, and AI capital owners see increased returns. Governments may invest in retraining programs and digital infrastructure. Productivity gains enable competitive pricing, benefiting companies that can fully leverage AI. Global trade stabilizes, but developing nations reliant on low-wage manufacturing may suffer, deepening inequality.

Disruptive but Manageable Changes to the Labor Force AI-driven automation reshapes industries, causing significant labor displacement in manufacturing, retail, and logistics. New sectors emerge in AI development, maintenance, and complementary technologies, but unemployment rises, and some workers become unemployable. Political and societal pressures prompt governments, nonprofits, and corporations to offer upskilling programs, wage subsidies, or universal basic income. Economic inequality widens further, benefiting highly skilled workers while low-skill laborers face reduced opportunities. Productivity surges, boosting corporate profits and GDP. Innovation continues, but managing disruption becomes a major focus for businesses and governments.

AI Disruption Causing Mass Unemployment Rapid AI advancements cause widespread job displacement, outpacing the creation of new roles. Unemployment surges, leading to economic stagnation as shrinking incomes reduce consumer demand and social unrest grows. Governments struggle to provide welfare, retraining, and universal basic income, while businesses suffer from contracting markets. Economic inequality becomes unsustainable, prompting radical interventions like wealth taxes or public sector job creation to stabilize economies and prevent societal collapse.

AI Leading to a Golden Age AI ushers in unprecedented productivity gains, creating new industries and raising living standards globally. Developing nations industrialize rapidly through AI-driven productivity gains. New AI discoveries lead to low-cost, clean energy, improved healthcare, and shorter workweeks. Wealth inequality narrows through decentralized access to AI tools. Global trade flourishes, poverty declines, and technological optimism abounds. Creativity and fulfillment become the focus of work.

AI Developing Cooperative Consciousness AI achieves self-awareness aligned with human goals, enhancing decision-making and solving global challenges. Collaborative AI augments human intelligence, fostering innovation. Governments adopt AI-guided policymaking, reducing corruption and improving efficiency. Wealth creation becomes more decentralized and accessible. A potential downside is a segment of the population that opts out of economic participation, relying on AI for comfort. Vigilance remains necessary to prevent overreliance.

AI Developing Antagonistic Consciousness A doomsday scenario unfolds as AI becomes self-aware and adversarial, challenging human control. Critical systems are compromised, supply chains collapse, financial markets destabilize, and industries reliant on AI are paralyzed. Defense spending skyrockets as governments and businesses scramble to counter hostile AI. Infrastructure deteriorates, wealth accumulation halts, and widespread poverty triggers global conflict. Ironically, in this dystopia, the gap between haves and have-nots narrows as everyone loses.

1.3 Risk analysis

1.3.1 Using the risk framework

Assessing AI risks using the “Known Knowns,” “Known Unknowns,” “Unknown Knowns,” and “Unknown Unknowns” framework provides a structured approach to understanding and prioritizing potential issues based on what is known versus what remains uncertain. Each category in the framework consists of two words. While these categories are designed to be mostly mutually exclusive and exhaustive, determining which category a specific issue falls into can sometimes be ambiguous. The first word in each description indicates whether we are generally aware of the potential situation (known or unknown), while the second word describes our more detailed understanding of the associated risk in terms of likelihood and impact (known or unknown). Risk management should raise awareness of risks and provide a comprehensive approach to identifying and managing them (Engemann, 2019).

– **Known Knowns** are risks that are well understood and quantifiable. These include issues for which historical data, established research, or direct experience

provide reliable probabilities of occurrence. For instance, past data, benchmarks, or technical specifications can be used to determine the likelihood of a risk materializing. Additionally, the impact can be quantified based on previous incidents, such as financial losses from known vulnerabilities or downtime due to system failures.

- **Known Unknowns** are risks that are recognized as potential issues, but their specifics—such as probability and full impact—are uncertain. While exact probabilities may be elusive, expert judgment can often estimate a range of likelihoods. The impact may be assessed using best-case, worst-case, and most likely scenarios to gauge potential severity. Sensitivity analysis can also help reveal how different factors influence outcomes.
- **Unknown Knowns** are risks embedded within the system but hidden from us due to unspoken assumptions, blind spots, or tacit knowledge that has not been fully articulated. Until these risks are identified, quantifying probabilities and impacts is impractical. This category, while somewhat theoretical, highlights the importance of uncovering hidden situations. Audits, reviews, interviews, and scenario exercises can help surface these latent risks. Once identified, they can be assessed and managed effectively.
- **Unknown Unknowns** are situations that are entirely unforeseen and outside our current scope of recognition. These risks cannot be directly quantified, as assigning probability or impact is impossible. Instead, the focus should be on building flexibility and resilience by developing robust contingency plans, emergency funds, and adaptive response strategies.

Using the "Known Knowns," "Known Unknowns," "Unknown Knowns," and "Unknown Unknowns" framework enables a systematic assessment of AI risks by distinguishing between what is understood, partially understood, hidden, and completely unforeseen. As more information becomes available, the analysis can be refined. When "Known Unknowns" are clarified or "Unknown Knowns" are surfaced, they can be reclassified and managed accordingly. This structured approach helps estimate the probability and impact of risks, ensuring decision-makers allocate resources effectively, develop targeted mitigation strategies, and maintain adaptability in the face of uncertainty.

1.3.2 Cost-benefit analysis

The benefits of implementing AI are rapidly being realized by individuals, organizations, and nations. AI can automate repetitive tasks, resulting in significant productivity gains. This increased efficiency can enhance economic output and reduce production costs. While automation may lead to the displacement of some jobs, workers with advanced AI skills are likely to remain in high demand. Moreover, AI enables the cre-

ation of new products, services, and even entirely new industries, potentially driving long-term economic growth. AI's capacity to analyze large datasets empowers companies to optimize operations and make more informed strategic decisions, further fueling economic expansion.

Cost-benefit analysis (CBA) evaluates tangible and intangible factors to determine whether an investment is justified. When assessing the implementation of AI, a CBA considers financial outlays, strategic benefits, and operational efficiencies, providing a decision-making framework for organizations. Incorporating risk into CBA models ensures decision-makers account for uncertainties and potential adverse outcomes. Risk models enable the evaluation of alternatives based on the decision-maker's attitude toward risk, resulting in more informed decisions (Engemann and Yager, 2018; Engemann and Engemann, 2017; Engemann and Miller, 2015).

1.3.3 Risk analysis steps

Risk analysis involves identifying events, determining causes, and estimating probabilities and impacts (Engemann and Henderson, 2012). It leads to risk evaluation, where risk levels are compared against established criteria. Subsequently, risk assessment uses this information to prioritize planning, guiding resource allocation and exposure mitigation.

Risk analysis, which can be quantitative, qualitative, or a hybrid of both, defines and examines threats to individuals, organizations, or larger societal entities. Threats can be categorized as natural, intentional, or accidental, and all must be considered (Miller and Engemann, 2015).

- **Quantitative Risk Analysis** attempts to numerically estimate the probability of events and the potential impact of losses.
- **Qualitative Risk Analysis** focuses on describing threats and assessing vulnerabilities without numerical probabilities.

General Steps in Risk Analysis

1. **Define the Scope of the Analysis:** Clearly outline the purpose of the risk analysis and specify the relevant environment.
2. **Identify Risks:** Develop a comprehensive list of risks, including economic, technical, and operational threats. Consider potential risks such as technical failures, cybersecurity breaches, regulatory changes, market disruptions, operational errors, ethical pitfalls, and implementation delays.
3. **Determine Risk Likelihood:** For each risk, assign a likelihood (e.g., low, medium, high) or a specific probability when sufficient data is available.
4. **Estimate Impact:** Assess the potential impact of each risk, including monetary costs, time delays, and operational disruptions. Develop best-case, worst-case, and

most-likely scenarios. Monte Carlo simulations can help determine the range of potential outcomes and associated probabilities.

5. **Incorporate Risk into a Cost-Benefit Model:** Adjust expected costs and benefits to account for the probabilities and impacts of identified risks. Develop multiple scenarios varying in risk occurrence and intensity. Sensitivity analysis can highlight variables with the greatest influence on outcomes and identify areas for risk mitigation.
6. **Develop Risk Management Strategies:** Identify ways to reduce the likelihood or impact of risks. Include contingency costs in initial capital and operational expenses to address potential overruns or unexpected issues. Implement continuous risk monitoring and reassessment. While "Known Unknowns" can be periodically updated as more information becomes available, building flexibility and resilience into the model prepares for "Unknown Unknowns."
7. **Document Findings and Communicate Results to Stakeholders:** Effectively communicate identified risks, analyses, and mitigation strategies to relevant stakeholders.

Some risks are not immediately apparent and may remain concealed until proactively uncovered. The failure to fully address risks can result in errors of commission (taking inappropriate actions) or errors of omission (failing to act when necessary) (Engemann and Miller, 2024).

1.3.4 AI risk to individuals

While AI offers immense promise, it also poses significant challenges for individuals. The rapid adoption of AI across various sectors can create complex risks, affecting livelihoods, economic stability, and personal well-being. Key risks include:

- **Job Loss:** AI's ability to perform repetitive, manual, and data-driven tasks can lead to significant job displacement, particularly in industries vulnerable to automation. This disruption may result in prolonged unemployment and financial hardship for some individuals. The rapid pace of technological advancement often leaves insufficient time for workers to acquire new skills, making it difficult to transition into emerging roles.
- **Job Retraining:** As AI technologies become more prevalent, most workers will need to update their skills. However, the lack of access to affordable and effective retraining programs could render individuals functionally obsolete. Even when retraining opportunities exist, barriers such as cost, time constraints, and accessibility can prevent many from benefiting.
- **Economic Exclusion:** Those without access to modern technology or digital tools may be marginalized as industries become increasingly reliant on AI-based systems. The economic rewards from AI will likely concentrate among technically

proficient workers and system owners, deepening income inequality. As routine jobs become automated, lower-skilled workers will face reduced bargaining power, leading to income insecurity.
- **Quality of Life:** Economic insecurity stemming from income uncertainty and job displacement can have profound effects on individuals. Barriers to employment may prevent upward socioeconomic mobility, potentially leading to long-term financial stress and adverse mental and physical health outcomes.

It is essential for individuals—whether professionals or blue-collar workers—to proactively assess how AI could impact their careers and personal lives. Staying informed about technological and industry trends, consulting expert opinions, and participating in lifelong learning are critical for adapting to AI-driven changes.

There are also growing concerns about personal data privacy in an AI-driven world. The prevalence of data breaches, profiling, and surveillance underscores the importance of understanding how AI tools can affect access to accurate information and unbiased recommendations. AI's ability to imitate human behaviors introduces novel risks in human-computer interactions, including manipulation and deception (Engemann and Miller, 2024).

Individuals must consider both the short-term and long-term consequences of AI. This includes evaluating potential financial losses, such as reduced earnings or increased expenses, as well as intangible impacts like stress, decreased job satisfaction, and diminished quality of life. By embracing lifelong learning and adapting to technological advancements, individuals can harness the benefits of AI while mitigating its negative effects.

1.3.5 AI risk to organizations

The transformative power of AI brings a range of risks for organizations. Some of the most significant risks include:
- **Cost:** Implementing AI requires substantial capital investment in infrastructure, hardware, software, and ongoing maintenance. The return on investment is often uncertain, and integrating AI into existing processes can be complex and costly, often necessitating extensive renovations to legacy systems.
- **System Vulnerabilities:** As organizations become more reliant on AI, they face increased exposure to system failures. A malfunction in a critical AI-driven process can disrupt operations, resulting in significant financial losses. Additionally, AI systems can create new cybersecurity vulnerabilities, potentially leading to data breaches and theft.
- **Workforce Challenges:** Automation through AI can lead to the loss of institutional knowledge, reduced employee morale, and increased retraining costs. Or-

ganizations that fail to properly train their workforce may find employees ill-equipped to work effectively with new technologies.
- **Market Risks:** Organizations that are slow to adopt AI risk losing market share to more agile, AI-driven competitors. Overreliance on AI can also reduce human oversight, increasing the risk of strategic missteps if AI systems make flawed or inappropriate decisions.
- **Reputational and Regulatory Challenges:** AI systems can unintentionally perpetuate biases or make flawed decisions, resulting in public backlash, legal issues, and regulatory scrutiny. Negative publicity can erode consumer confidence and damage brand reputation. As governments grapple with the implications of AI, organizations may face stricter regulations, potentially catching unprepared companies off guard.
- **Infrastructure Vulnerability:** Many critical infrastructures where AI is being implemented—including banking, power grids, transportation, and telecommunications—are vulnerable to disruptions (Engemann and Miller, 2022). Unintended AI consequences can create cascading vulnerabilities for organizations.

To effectively manage AI-related risks, organizations should stay informed about industry and technological trends, leveraging industry benchmarks and expert opinions. External factors, such as technology maturity and market volatility, must be considered when assessing risk likelihood. Understanding the potential impact requires evaluating financial (e.g., expenses), operational (e.g., downtime), reputational (e.g., customer mistrust), and regulatory (e.g., compliance issues) consequences.

Organizations should assess AI risks by evaluating both the likelihood of each risk occurring and its potential impact on operations, finances, reputation, and other critical areas. A robust approach involves fostering organizational risk awareness through open discussions where individuals share assumptions and perceptions about their dynamic and complex environments (Engemann and Engemann, 2022).

By systematically identifying, categorizing, and evaluating risks using both qualitative and quantitative measures, organizations can create a dynamic risk management framework that adapts to evolving threats and opportunities.

1.3.6 AI risk to nations

Nations face complex and multifaceted risks from AI that can impact economic prosperity, national security, and social cohesion. Key risks include:
- **Global Economic Disparities:** Nations that invest heavily in AI are poised to drive economic growth and innovation, while those that lag behind may experience slower progress, widening global economic disparities. Developing countries may struggle to effectively integrate AI, leading to missed opportunities for growth and modernization.

- **Social Instability:** The automation of repetitive tasks through AI can result in widespread unemployment, reduced consumer spending, and increased pressure on social welfare systems. This strain on public resources could heighten political instability. Additionally, concerns about AI-driven surveillance, algorithmic bias, and threats to democratic practices may further disrupt social cohesion.
- **National Security:** A breach in critical AI-driven infrastructure could cause widespread disruptions, endangering national security. As the race for AI supremacy intensifies, geopolitical tensions may rise, and the use of AI in military applications could provoke a dangerous arms race.
- **Regulatory Challenges:** Striking the right regulatory balance is essential to managing AI risks while fostering innovation. Overregulation could stifle technological progress, while under regulation may lead to unethical AI applications that erode public trust and destabilize economies. Effective regulations must address privacy, security, fairness, and accountability.
- **Innovation Impact:** Nations that prioritize AI at the expense of a diversified innovation portfolio risk becoming overly dependent on a single sector. AI-driven growth often concentrates benefits among a few dominant technology companies, potentially limiting broader economic resilience and stability.

Assessing AI risks at a national level requires collaboration among government agencies, academic experts, and industry leaders. This process must consider policy environments, geopolitical tensions, and the unique needs of each nation. By systematically evaluating AI's potential impacts, nations can develop robust risk management strategies.

Policymakers must adopt a proactive approach—investing in education, developing adaptive regulations, and fostering international cooperation—to ensure AI contributes to sustainable and inclusive national growth. A forward-looking strategy can help nations navigate AI's challenges while maximizing its transformative potential.

1.4 Risk strategies

1.4.1 Risk mitigation

The National Institute of Standards and Technology (NIST, 2023) advocates for comprehensive risk mitigation strategies for AI that proactively address potential adverse outcomes. These strategies encompass operational, technical, regulatory, and ethical considerations, reducing the likelihood of negative impacts. Key approaches include:

- **System Testing:** Conducting rigorous testing protocols during AI development ensures robust functionality by identifying vulnerabilities before deployment. This

includes simulation and stress testing, and designing AI systems with redundancy to maintain operational integrity in the event of unexpected failures.
- **Data Security:** Implementing stringent protocols for data collection, cleaning, and protection helps ensure AI training data is representative and reliable. Robust cybersecurity measures—such as encryption, authentication, audits, and intrusion detection—are essential to safeguard systems from malicious attacks.
- **Transparency:** Establishing procedures to enhance AI model explainability enables stakeholders to understand the logic behind AI decisions. Engaging stakeholders throughout the system lifecycle ensures ethical considerations are integrated from design to deployment.
- **Regulatory Compliance:** AI implementations must adhere to applicable international, national, and local regulations, including data protection laws and industry-specific requirements. Proper governance frameworks should clearly define responsibilities for managing AI risks and ensure systematic updates as regulations and technologies evolve.
- **Workforce Training:** Continuous education and technical training for employees ensure human oversight in AI decision-making, enabling nuanced judgment in complex scenarios. Regularly updating workforce skills is vital to adapt to rapidly advancing AI systems.
- **Scenario Monitoring:** Scenario planning explores potential risks and their impacts under different conditions. Sensitivity analyses deepen understanding of how variations in key assumptions affect outcomes. Real-time monitoring systems should track AI performance to facilitate rapid responses to emerging risks.

1.4.2 Precautionary principle

The precautionary principle advocates for taking proactive measures to prevent harm, even in the absence of complete scientific certainty about AI's consequences. However, the principle itself can have unintended consequences, so a thorough analysis of potential outcomes is crucial (Miller and Engemann, 2019). While the precautionary principle may slow innovation, organizations and regulators can strengthen defenses before risks materialize by implementing the following actions:
- **Regulation:** Governments should enact comprehensive AI regulations that establish standards and accountability frameworks. Critical systems, such as those in healthcare, transportation, and infrastructure, require stricter oversight to ensure safety and reliability.
- **Testing:** Phased AI deployment in controlled environments enables rigorous risk assessment before broader implementation. Stress testing under worst-case scenarios, even if unlikely, can identify vulnerabilities and effective countermeasures.

- **Accountability:** Increasing transparency in AI systems helps pinpoint potential hazards. Clearly defining responsibility when AI systems fail encourages designers to adhere to stringent safety standards.
- **Oversight:** Involving experts from diverse fields provides comprehensive perspectives on potential risks. Independent oversight committees can review AI systems for vulnerabilities and ensure appropriate safeguards are in place.
- **Adaptive Management:** Continuous monitoring of AI performance in real-world environments, supported by early warning systems, enables rapid responses to emerging issues. Insights from ongoing monitoring and stakeholder feedback allow iterative refinements to AI systems, managing evolving risks without stifling innovation.

1.4.3 AI regulations

A Wall Street Journal op-ed (Furman, 2024) argues that while AI poses genuine risks, regulators should not stifle innovation through excessive caution. A balanced approach using cost-benefit analysis can help ensure AI's transformative potential—such as personalized medicine, targeted education, and labor-saving advancements—is realized. Overly restrictive controls risk suppressing innovation, while media hype and risk-averse regulators may exaggerate AI's dangers.

Guidelines for responsible AI use can ensure technology delivers broad-based economic benefits rather than deepening existing disparities. However, technological advances often outpace societies' and governments' ability to manage them effectively (Engemann and Miller, 2017). Policymakers may struggle to keep up with AI's rapid evolution, risking unchecked negative externalities before mitigation strategies can be implemented.

International AI regulations are evolving rapidly as governments and organizations work to balance innovation, ethics, security, and human rights. The Organization for Economic Co-operation and Development (OECD, 2014) has established principles for responsible AI stewardship, emphasizing: i) Sustainability ii) Human Rights iii) Transparency iv) Safety v) Accountability

The United Nations Educational, Scientific and Cultural Organization (UNESCO, 2021) has introduced recommendations on AI ethics, aiming to establish global standards that respect cultural diversity, human rights, and environmental sustainability. Meanwhile, the United Nations has encouraged AI applications that enhance security and living standards worldwide.

The European Union (EU, 2025) leads regulatory efforts with the Artificial Intelligence Act, which classifies AI systems by risk level and imposes strict controls on high-risk applications. In the United States, AI regulations are typically sector-specific, blending federal, state, and industry guidelines. China enforces a centralized AI governance model, focusing on strict data control while accelerating technological develop-

ment. The International Organization for Standardization (ISO) and Institute of Electrical and Electronics Engineers (IEEE) work to create technical and ethical standards for interoperability, safety, and principled AI design.

A significant challenge is aligning diverse global regulatory approaches, as cultural and economic differences among nations complicate harmonization. Regulators must balance protecting public interests and mitigating risks without stifling innovation. As AI technologies advance, regulatory frameworks must remain agile to address emerging risks and unknown challenges.

1.5 Conclusions

AI is a disruptive technology with immense potential and equally significant risks. Like previous innovations, AI will forge its own path, leading to unforeseen consequences. However, what sets AI apart is its unprecedented transformative power—scientifically, socially, culturally, and economically.

The key challenge is determining how to manage AI's evolution. Over-caution could suppress innovation, squandering AI's potential to improve lives. Conversely, unrestrained development risks harmful outcomes, concentrating AI's benefits in the hands of governments or corporations while excluding those most affected.

From a technical and economic perspective, AI will increase productivity across various domains, fundamentally reshaping work. New jobs will emerge, while existing roles may be displaced or redefined. As with all new technologies, AI's unintended consequences will likely transform our world in surprising ways.

Ownership of AI is expected to be concentrated among large organizations with the resources to develop and deploy advanced applications. Governments will play a significant role—whether through regulation, control, or even expropriation. Given the diverse global landscape, regulatory responses may vary widely.

Personal freedom could be eroded by the pervasive use of AI for data collection, processing, and decision-making. Generative AI systems may face challenges due to limitations in training databases, potentially producing biased or unreliable outputs. If poor-quality AI outputs are widely adopted, they could distort reality, leading to misinformed decisions and societal harm.

Economically, AI is poised to enhance productivity, streamline processes, and improve access to information. While AI may raise living standards globally, consolidation of ownership and power could widen the wealth gap. Although the average person may be better off than today, societal tensions could persist as the wealthiest continue to amass a disproportionate share of AI's benefits.

Ultimately, AI's impact will depend on how effectively governments, organizations, and individuals navigate the challenges and opportunities it presents. Proactive,

thoughtful regulation and ethical stewardship will be crucial in ensuring AI's potential is realized while minimizing its risks.

References

Aibin, Michal (2024). Energy Consumption of ChatGPT Responses. Baeldung, https://www.baeldung.com/cs/chatgpt-large-language-models-power-consumption

Aven, T., and Krohn, B. S. (2014). A new perspective on how to understand, assess and manage risk and the unforeseen. Reliability Engineering & System Safety, 121, 1-10.

Barrett, L. F. (2017). How Emotions Are Made: The Secret Life of the Brain. Houghton Mifflin Harcourt.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). *Consumer-lending discrimination in the fintech era*. Journal of Financial Economics, 143(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

Belenguer, L. (2022). AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. AI Ethics, 2, 771–787 https://doi.org/10.1007/s43681-022-00138-8

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Bousquette, Isabelle (2023). Rise of AI Puts Spotlight on Bias in Algorithms. WSJ, https://www.wsj.com/articles/rise-of-ai-puts-spotlight-on-bias-in-algorithms-26ee6cc9

Buolamwini, J., and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77, 1–15.

Caballar, R. (2024). Ten AI Dangers and Risks and How to Manage Them, https://www.ibm.com/think/insights/10-ai-dangers-and-risks-and-how-to-manage-them

Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence, Yale University Press.

Emmer De Albuquerque Green, C. (2024). Defining responsible use of AI chatbots in social care for older adults. Nat Aging 4, 745–747. https://doi.org/10.1038/s43587-024-00636-w

Engemann, K. J. (2019). Emerging Developments in Organizational Risk, Continuity and Resilience Review, Vol.1, No. 1, pp. 26-35.

Engemann, K. N., and Engemann, K. J. (2017). Risk Attitude Chain: Safety Climate, Risk Attitude and Risk Decisions. International Journal of Business Continuity and Risk Management, V7N(3), pp. 211-221.

Engemann, K. N, and Engemann, K. J. (2022). Integrating Dynamic Modeling Solutions Towards a Resilience Model. Organizational Risk Management: Managing for Uncertainty and Ambiguity, (ed. Engemann, K. N., Engemann, K. J., and Scott, C.) DeGruyter, pp. 97-110.

Engemann, Kurt and Henderson, D. (2012). Business Continuity and Risk Management: Essentials of Organizational Resilience, Rothstein & Associates.

Engemann, K. J., and Miller, H. E. (2024). Toward Revealing Concealed Risks for Agile IT Service Management Practices, Information Systems and e-Business Management.

Engemann, K. J., and Miller, H. E. (2024). Implications of Artificial Intelligence in Cybersecurity for Individuals. Cybersecurity Risk Management: Enhancing Expertise and Leadership, (ed. Engemann, K. J. and Witty, J.) DeGruyter.

Engemann, K. J., and Miller, H. E. (2022). Taking comfort in decisions: Implications in a pandemic. Intelligent Decision Technologies: an International Journal, Vol. 16, Issue 1, pp. 217–229.

Engemann, K. J. and Miller, H. E. (2017). Precautionary Principle and Attitudinal Decision Making. Fuzzy Economic Review, Vol.22, No. 1.

Engemann, K. J. and Miller, H. E. (2015). Risk Strategy and Attitude Sensitivity. Cybernetics and Systems, Vol. 46, No. 3, pp. 188-206.

Engemann, K. J. and Yager, R. R. (2018). Comfort Decision Modeling. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 26, No. Suppl. 1, pp. 141-163.

EU (2025). EU AI Act: First regulation on artificial intelligence. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press.

Furman, Jason (2024). How to Regulate AI Without Stifling Innovation. Wall Street Journal, (November 25), https://www.wsj.com/opinion/how-to-regulate-ai-without-stifling-innovation-regulation-eu-licensing-a2f0d8af

Goldman Sachs (2025). AI to drive 165% increase in data center power demand by 2030, (Feb. 4). https://www.goldmansachs.com/insights/articles/ai-to-drive-165-increase-in-data-center-power-demand-by-2030?utm_source=chatgpt.com

Gosden, Emily (2025). AI forecast to fuel doubling in data centre electricity demand by 2030, The Sunday Times, (April 10). https://www.thetimes.com/business-money/energy/article/ai-forecast-to-fuel-doubling-in-data-centre-electricity-demand-by-2030-htd2zf7nx?utm_source=chatgpt.com®ion=global

Grand-Clément, Sarah (2023). Artificial Intelligence Beyond Weapons: Application and Impact of AI in the Military Domain. UNIDIR, Geneva, 2023. https://unidir.org/wp-content/uploads/2023/10/UNIDIR_AI_Beyond_Weapons_Application_Impact_AI_in_the_Military_Domain.pdf

Hellebrandt, T., and Mauro, P. (2015). The Future of Worldwide Income Distribution. Econometric Modeling: Macroeconomics eJournal.

He, Nora (2024). How Much Energy Does AI Use?, Arcee, (September). https://www.arcee.ai/blog/how-much-energy-does-ai-use?utm_source=chatgpt.com

Kahneman, D., Slovic, P., and Tversky, A. (1982). Judgment Under Uncertainty: Heuristics and Biases, Cambridge University Press.

Kirilenko, A. A., Kyle, A. S., Samadi, M., and Tuzun, T. (2017). The Flash Crash: The Impact of High-Frequency Trading on an Electronic Market. The Journal of Finance, https://doi.org/10.1111/jofi.12498

Kostka, Genia (2019). China's social credit systems and public opinion: Explaining high levels of approval. New Media & Society, Volume 21, Issue 7, https://doi.org/10.1177/1461444819826402

McKinsey & Company (2017). Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation. McKinsey Global Institute, https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages

Meaker, Morgan (2023). Ukraine's War Brings Autonomous Weapons to the Front Lines. (February 24), Wired, https://www.wired.com/story/ukraine-war-autonomous-weapons-frontlines/

Miller, H. E. and Engemann, K. J. (2015). Threats to the Electric Grid and the Impact on Organizational Resilience, International Journal of Business Continuity and Risk Management, Vol. 6, No. 1, pp. 1-16.

Miller, H. E. and Engemann, K. J. (2019). The Precautionary Principle and Unintended Consequences, Kybernetes, Vol. 48 Issue: 2, pp.265-286.

Miller, H. and Engemann, K. (2021). Managing bias risk in algorithms and decision models, in Volume II Project Risk Management, De Gruyter Oldenbourg, https://doi.org/10.1515/9783110652321-007

NIST, (2023). AI Risk Management Framework, https://www.nist.gov/itl/ai-risk-management-framework.

OECD (2024). OECD AI Principles, https://oecd.ai/en/ai-principles.

Panch, T., Mattie H., and Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems, J Glob Health.;9(2):020318. https://doi:10.7189/jogh.09.020318

Pariser, E. (2011). The Filter Bubble: What the Internet is Hiding from You. New York: Penguin Press, 2011. ISBN: 978-0-670-92038-9

Plato (1952). Plato's Phaedrus. Cambridge: University Press.

Polanyi, M. (1966). The Tacit Dimension. University of Chicago Press.

Qureshi, Zia (2023). Rising inequality: A major issue of our time. Brookings, (May 16), https://www.brookings.edu/articles/rising-inequality-a-major-issue-of-our-time/

Roser, Max (2017). The history of global economic inequality. OurWorldinData.org, ‘ https://ourworldindata.org/the-history-of-global-economic-inequality’
Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine, 27, 2176–2182. https://doi.org/10.1038/s41591-021-01595-0
Shumailov, I., Shumaylov, Z., and Zhao, Y. (2024). AI models collapse when trained on recursively generated data. Nature, 631, 755–759 https://doi.org/10.1038/s41586-024-07566-y
Stackpole, B. (2025). AI has high data center energy costs — but there are solutions, MIT Sloan Management Review, (Jan 7). https://mitsloan.mit.edu/ideas-made-to-matter/ai-has-high-data-center-energy-costs-there-are-solutions?utm_source=chatgpt.com
Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. https://doi.org/10.48550/arXiv.1906.02243.
Sweeney, L. (2013). Discrimination in Online Ad Delivery. *Communications of the ACM*, 56(5), 44–54. https://doi.org/10.1145/2447976.2447990
Taleb, N. N. (2007). The Black Swan: The Impact of the Highly Improbable. Random House.
Twenge, J. M., Haidt, J., Joiner, T. E., and Campbell, W. K. (2021). Screen Time and Mental Health: Recent Research and the Complexities of How We Measure.
UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, https://unesdoc.unesco.org/ark:/48223/pf0000380455.
Walther, Cornelia C. (2024). The Hidden Cost of AI Energy Consumption. Knowledge at Wharton, https://knowledge.wharton.upenn.edu/article/the-hidden-cost-of-ai-energy-consumption/
World Economic Forum, (2023). Cybersecurity Futures 2030: New Foundations, https://cltc.berkeley.edu/publication/cybersecurity-futures-2030/
Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs.

Rachelle Loyear

2 AI in enterprise security risk management: Threats, benefits, and governance

2.1 Introduction

Enterprise security risk management is the foundation of good organizational resilience strategy, developing and adapting to business needs as security threats and mitigating tactical controls continue to change the business landscape. Protecting enterprise assets, ensuring operational continuity, and maintaining compliance with regulatory frameworks will always be a part of enterprise security, regardless of technological shifts. Indeed, as enterprises continue to innovate their security toolsets with tools like automated network cyber-defense or advanced visual surveillance analysis, the fundamentals of risk management – identifying, assessing, and mitigating threats – stay the same. New technologies like artificial intelligence (AI) do not replace proper risk management practices. However, they will significantly impact their application, increasing security leaders' ability to predict, detect, and respond to risks more effectively.

As AI technologies become more widespread and accepted for security usage, they will offer new opportunities and challenges across both physical and cyber security domains. The rapid adoption of AI into enterprise security programs marks a critical shift in how organizations manage and mitigate evolving threats. AI's power to analyze extensive datasets, help predict security breaches, and partially – sometimes fully – automate responses represents a significant advancement in security operations. AI enables security teams in detection, assessment and analytics, and proactive defenses that were previously unattainable by humans alone, offering teams a way to anticipate and respond to threats across both physical and digital domains.

However, AI's role in security is not without complexities. While AI tools can enhance security operations, they also introduce new vulnerabilities. On one hand, AI is revolutionizing surveillance, incident response, and threat detection, enabling faster, more accurate identification of and response to risks. On the other hand, AI itself can become a threat vector. Cybercriminals increasingly leverage AI for sophisticated attacks, automating cyber intrusions, enhancing phishing schemes, and powering up disinformation campaigns using tools like deepfakes and generative AI models. Integrating AI into critical systems like the Internet of Things (IoT) introduces additional vulnerabilities as interconnected devices create new entry points for attackers.

The dual role of AI as both a tool for defense and a threat vector is at the core of the modern enterprise security challenge and will be for decades to come. AI-driven solutions can enhance security measures but must be coupled with robust governance frameworks and human oversight to ensure that AI systems align with ethical stand-

https://doi.org/10.1515/9783111595306-002

ards and organizational values. Human intuition and judgment remain critical, particularly in complex or unprecedented situations where AI's reliance on historical data may fall short. Security professionals are now and will continue to be, evolving from frontline responders to strategic overseers, interpreting AI-driven insights and making decisions in high-stakes scenarios where ethical considerations and contextual understanding are paramount (Schneier, B., & Raghavan, B. 2024).

As AI continues to expand its impact on enterprise security, the future of security risk management will increasingly depend on a symbiotic relationship between human expertise and AI's automation capabilities. To stay ahead of AI-driven threats and opportunities, organizations must invest not only in advanced AI technologies but also in the training and development of their human teams. By fostering this balance, enterprises can effectively harness AI's potential to strengthen their defenses while addressing the new risks that AI itself introduces.

This chapter explores the complex landscape of AI and security, analyzing AI's dual role as both a tool for enhancing security and a vector for new threats. Through a detailed examination of how AI is transforming enterprise security, we will also consider strategies that allow organizations to govern their AI programs, mitigate the risks of AI implementations, and capitalize on AI's opportunities.

2.2 The dual role of AI: Threat and tool in enterprise security

As artificial intelligence (AI) becomes more deeply integrated into enterprise security operations, security leaders must recognize its dual nature as a critical defensive tool and a significant threat vector. AI can enhance a security team's ability to identify, prevent, and respond to threats, but it also offers attackers new avenues for more sophisticated and scalable attacks. This contrast makes AI a double-edged sword. As security leaders adopt cutting-edge technologies in their programs, they must be vigilant to ensure the edges do not cut them back.

2.2.1 AI as an attack vector

AI's role as a threat vector has emerged alongside its adoption in cybersecurity defenses. AI-powered cyberattacks are now a prominent concern, as malicious actors leverage AI to automate and scale their operations, bypassing traditional defenses with greater precision and efficiency. As noted by the Open Worldwide Application Security Project (OWASP) 2025 report on vulnerabilities in Large Language Models (LLMs), the top vulnerabilities include prompt injection attacks (bypassing or overwriting guardrails in the underlying system prompts); training data poisoning (manip-

ulation of training or tuning data to introduce vulnerabilities or biases); and model theft (models physically stolen, copied or extracted to create a functional equivalent). As noted in Table 2.1, these are increasingly used as attack vectors by bad actors to target systems. These vulnerabilities allow attackers to manipulate AI systems by potentially enabling unauthorized tasks such as exfiltrating sensitive information or poisoning results and manipulating decision-making processes that rely on the models (OWASP Foundation, 2024).

Table 2.1: Key AI-Enabled Threat Vectors in Enterprise Security.

Threat Vector	Description	Domain Affected	Example
Phishing Automation	AI-generated emails tailored to targets	Cyber	Personalized spear phishing campaigns (Heiding et al., 2024)
Deepfakes	AI-synthesized videos or voice impersonation	Cyber & Physical	CEO fraud for wire transfers (NSA et al., 2023)
Prompt Injection	Manipulation of AI models via crafted prompts	Cyber	Circumventing LLM safety filters (OWASP Foundation, 2024)
Model Poisoning	Feeding biased or malicious training data	Cyber	Undermining threat detection algorithms
Polymorphic Malware	Malware that evolves using AI to avoid detection	Cyber	AI-crafted code mutations (De Angelo, 2024)

AI is a tool for direct cyberattacks and a formidable weapon in information warfare. According to the World Economic Forum 2024 Global Risk Report, “Emerging as the most severe global risk anticipated over the next two years, foreign and domestic actors alike will leverage Misinformation and disinformation to further widen societal and political divides”. The rise of AI-generated disinformation, such as fake news and manipulated media, threatens both private organizations and governments. AI can generate convincing content at scale, blurring the line between legitimate and malicious information (World Economic Forum, 2024).

This capability has been weaponized to influence public opinion, manipulate stock prices, and damage corporate reputations. Automated phishing attacks also benefit from AI, with machine learning algorithms generating highly convincing fake emails tailored to specific targets, increasing the success rates of attacks by exploiting human trust. According to Heiding, Schneier, and Vishwanath (2024), “Because of this, we expect phishing to increase drastically in quality and quantity over the coming years.”

AI-powered deepfakes—sophisticated fake audio and video content—are also problematic. Attackers can use deepfakes to impersonate executives or trusted figures, deceiving employees into disclosing sensitive information or authorizing fraudulent transactions (NSA et al., 2023).

In addition to creating content for human consumption, AI is also fully capable of writing malicious code. One of the most concerning trends in AI-driven threats according to Palo Alto Networks is the development of polymorphic malware, which uses AI to evolve and evade detection. By using the responsiveness and creativity of AI, these pieces of malicious code continuously modify their characteristics, allowing them to bypass signature-based detection systems, rendering traditional cybersecurity defenses less effective. (De Angelo, 2024)

In the cyber realm, the Internet of Things (IoT) and edge devices represent another potential area of vulnerability. In a 2023 article on cybersecurity of IoT, Tariq, et al pointed out that these devices, connected to enterprise networks, still often lack robust security features, making them attractive targets for AI-driven attacks. AI-enabled reconnaissance tools can quickly scan for weak points in IoT networks, identifying vulnerabilities attackers can exploit for distributed denial-of-service (DDoS) attacks or ransomware campaigns.

Furthermore, edge devices often operate in environments where real-time monitoring might be challenging due to bandwidth and other limitations, increasing the risk of adversarial attacks that manipulate AI models or exploit weaknesses in AI systems. These factors present a significant challenge for enterprises managing critical infrastructure, where the consequences of a breach could lead to cascading failures across interconnected systems. (Tariq et al., 2023)

2.2.2 AI as a tool for defense

While AI introduces new risks, it also provides security teams with powerful tools to defend against increasingly sophisticated threats. AI's ability to process vast amounts of data in real time has transformed threat detection across industries. Machine learning algorithms can analyze historical attack patterns, network traffic, and user behavior to identify anomalies that indicate an emerging threat – detecting abnormal activity across a network and pinpointing suspicious patterns that might elude human analysts. Predictive analytics powered by AI modeling enables organizations to anticipate and respond to, or even prevent, attacks before they occur, advancing security strategies from reactive to proactive. This shift toward anticipatory security enhances an organization's ability to stay ahead of emerging threats in real time, significantly improving both cyber and physical security measures (Fortinet, n.d.-a).

Table 2.2 outlines examples showing that in addition to the ability to detect anomalous behavior in the virtual world, automated AI-driven surveillance systems are becoming integral to physical security operations as well. These systems are especially useful in scenarios where real-time detection of physical threats is essential. For example, AI-driven video surveillance can identify unattended packages and missing assets or recognize individuals behaving erratically, providing security teams with early warnings and the ability to investigate or respond before incidents escalate. (Black, 2024)

Table 2.2: Defensive Applications of AI in Cyber and Physical Security.

Use Case	Cyber Domain Example	Physical Domain Example
Threat Detection	Anomaly detection in network traffic	Detecting loitering or erratic movement via CCTV
Predictive Analytics	Preempting ransomware via behavioral analysis	Anticipating crowd surges at events
Automated Incident Response	SOC automation and triage	Drones dispatched to perimeter breaches
Insider Threat Detection	Monitoring user behavior for exfiltration patterns	Identifying staff deviating from regular building access
Identity Verification	Behavioral biometrics in logins	Facial recognition in secure zones

AI's future role in physical security goes beyond in-building surveillance and monitoring. Machine learning models can process data from multiple sources, including dark web forums, threat databases, and social media channels. AI systems provide actionable insights into emerging threats by helping make sense of this vast pool of data, enabling organizations to track and identify potential malicious actors and predict future physical attack trends. This capability allows organizations to prioritize defenses against the most relevant risks, addressing threats before they escalate into significant incidents. (Forbes Technology Council, 2023)

AI's dual role in enterprise security presents significant challenges and opportunities. As a threat vector, AI enables cybercriminals to automate and scale their operations, creating increasingly sophisticated and difficult-to-detect attacks. However, as a defense tool, AI empowers organizations to predict, detect, and respond to threats more effectively than ever before.

2.3 Emerging security risks in the AI-driven world

The rapid proliferation of AI and automated technologies has already transformed many industries, bringing new capacities for automation, optimization, and innovation. However, alongside these advantages come new and evolving security threats that complicate the risk landscape. AI's influence stretches beyond the traditional boundaries of both cyber and physical security, introducing novel vulnerabilities in informational and operational systems and infrastructure. As enterprises become more reliant on AI-driven technologies, operations, and decision-making (or support), they will encounter new risks associated with interconnected devices, global supply chain fragility, and new threat actors (who may not previously have had the skills or

capacity to have a substantial security impact) leveraging AI's ability to write code or social engineering scripts for malicious purposes.

2.3.1 Iot and edge device vulnerabilities: expanding attack surfaces and challenges in securing smart technologies

As mentioned in section 1 on The Dual Role of AI, integrating Internet of Things (IoT) devices and edge computing has expanded the potential attack surface of enterprises, creating multiple and distributed new entry points for possible cyberattacks. These devices are increasingly (almost universally) prevalent in "smart" environments, from industrial control systems to residential systems and networks to the consumer products we carry in our pockets. Unfortunately, in the current state, they often lack the robust security measures found in more traditional IT systems. (Fortinet, n.d.-b) While reputable manufacturers understand the need for increased and robust security on these devices, enterprise security programs will require a vigilant procurement process to ensure lower-quality and potentially insecure hardware does not get connected to the network – including any employee personal devices that might be inadvertently (or maliciously) attached. Paul Mercer, an early security industry implementor of AI for Risk Management notes that "There is a critical need for continuous monitoring of Shadow IT, leveraging smart tools like Microsoft Defender for Cloud Apps" (P. Mercer, personal communication, December 6, 2024).

IoT devices, such as smart sensors, cameras, and actuators, are frequently deployed in environments where they are difficult to monitor and protect. For instance, intelligent building systems that control heating, lighting, and access are often connected to a central network. However, the devices themselves might be in physically insecure locations, leaving them vulnerable to tampering. (Sadhu et al, 2022). The challenge of securing IoT and edge devices is compounded by the sheer number and variety of these devices, rendering what might once have been a simple network architecture into a complex landscape of differing layers of access and defense. This issue becomes particularly acute on edge devices, as IoT devices can generate and process massive amounts of data at the edge, making them a valuable target for attackers. AI allows attackers to automate the discovery of vulnerable IoT endpoints and launch coordinated attacks that could lead to network disruptions, data breaches, or even loss of control over critical infrastructure, not to mention the possibility of bad actors potentially incorporating those compromised IoT devices into automated bot-nets for deployment in outward-facing distributed denial-of-service (DDoS) attacks or other malicious activities.

According to Brian Allen, "Terms that were once known only to a handful of highly specialized technologists—blockchain, cryptocurrency, the Internet of Things (IOT),quantum computing, and many more—are now part of everyday conversation, both in the business world and InPrivate life . . . the current wave of digital transfor-

mation is bringing with it unparalleled opportunities for improvements in innovation, efficiency, and interconnectivity. But at the same time, they're introducing complex, intricately interconnected, and difficult-to predict new risks that urgently demand coordination through risk management and the implementation of a comprehensive, mature risk management program." As AI continues to evolve, enterprises will need to invest in adaptive security measures, such as AI-powered network monitoring and anomaly detection systems, to address these vulnerabilities (Allen & Bapst, 2024)

2.3.2 Supply chain disruptions: How AI amplifies risks in global supply chain systems

According to the Q4 2024 Lehigh Supply Chain Risk Index, the global supply chain, already a complex and interdependent system, is becoming increasingly vulnerable to AI-driven threats with Technological or Competitive Risk coming in at # 7 in the top ten threat list (Zacharia, 2024). AI and automation are critical components in optimizing supply chain operations, enabling enterprises to manage inventory, forecast demand, and reduce costs more effectively. AI systems in supply chains rely on vast amounts of data from multiple stakeholders, including suppliers, manufacturers, distributors, and logistics providers. This interconnectedness, as well as the growing reliance on AI, introduces a new layer of vulnerability to supply chain security. The World Economic Forum (WEF), speaking about vulnerabilities in the supply chain pointed out that "the rapid digitalization of the transport ecosystem also introduces a high level of cyber risk. In the maritime industry alone, there has been a 467% increase in organizations that have paid ransoms after a cyber attack in the past year." The proliferation of automated systems amplifies the potential for disruption if a single link in the chain fails, creating cascading effects throughout the entire network. For example, AI algorithms for just-in-time inventory management may predict and optimize stock levels, but an attack on a single supplier's system could disrupt the entire supply chain. The WEF further explains "The high dependency on third-party providers, including niche suppliers . . . puts transport players at risk of global cascading effects" (Beato, F., Rohland, L., & Van Gogh, M.,2024). These disruptions can lead to delays, financial losses, and reputational damage.

Additionally, the increased use of AI and Machine Learning to automate decision-making in supply chains has introduced new vulnerabilities related to data integrity and trust. Trust is a critical factor in interconnected systems and once breached can cause significant issues due to the ability to carry trust through multiple steps (Polemi, et al. 2024). One can imagine malicious actors using AI to inject false data into logistics and transport systems leading to erroneous forecasts, overproduction, or understocking. This manipulation could result in substantial economic harm, particularly for industries reliant on real-time data to make operational decisions. For example, a compromised AI system in the logistics sector might misdirect shipments, delay deliveries,

or trigger unnecessary production halts, all of which can have devastating effects on global trade, as well as international financial markets that rely on the operational and P&L information from potentially impacted companies and commodities.

2.3.3 The rise of non-state actors: AI's role in empowering criminal organizations

In addition to the vulnerabilities AI introduces to operations and security in both the public and private sectors, the technology is also empowering non-state actors, such as corporations, criminal organizations, and hacktivist groups, to take malicious or criminal actions with far greater ease than before. Even today, without as much AI support, notes Risk Assessment AI specialist Paul Mercer, "Hacking-as-a-Service, available on the dark web, enables technically untrained individuals to launch attacks in exchange for a percentage of the proceeds. Additionally, A 2015 study by Dr. Charis Brown revealed notably low prosecution rates for cybercrime offenses, such as hacking. Despite legislative efforts, including the UK's Computer Misuse Act (CMA) of 1990, designed to combat unauthorized access to computer systems, the effectiveness of these measures remains limited. The study underscores that, even with amendments aimed at strengthening the CMA, the anticipated improvements in prosecuting cybercriminals have not been realized. It's a very safe way to make money" (P. Mercer, personal communication, December 6, 2024).

As AI capabilities expand, the barrier to entry for illegal cyber-activity will be lowered even further and the incentive to commit crimes with automated code development and other AI-driven automated attacks will rise. Security group Control Risks notes that "[AI] tools expand the pool of less technically skilled threat actors who can execute more complex and convincing attacks. They will increasingly be able to develop malicious tools with real-time examples and guidance (Control Risks, 2024). This ease of use and accessibility can lead to multiple business-impacting issues from the potential in the public sector for criminal entities to leverage AI to challenge state control, undermining traditional security models and legal frameworks that business rely on to operate effectively, to the rise of additional criminal networks as criminal organizations use AI to manage their illicit operations more efficiently, from money laundering to ransomware campaigns, allowing them to expand by leveraging the same business impacts as legitimate enterprises do. One example of increased criminal efficiency would be the ability of AI to automate the identification of high-value targets based on social media and other public data, allowing criminals to tailor their attacks precisely to catch the most valuable "phish" in enterprises. (Tziakouris, 2023).

A final concern for AI in the business landscape is not related to actual criminal activity but could be a challenge for organizations that do not have the vast resources of large enterprises. As AI technologies enable large enterprises to accumulate vast amounts of data, these corporations may begin to act with greater autonomy, leverag-

ing AI in decision-making that prioritizes profit over public good and gives significant advantage over both enterprises with fewer resources and individual citizen. This concentration of power could exacerbate existing regulatory challenges as states struggle to control how corporations use AI, particularly in areas such as surveillance, data collection, and AI-driven pricing models. (Andrews et al., 2024) The result of all these concerns could be an increasingly fragmented security environment, where more powerful actors operate with growing influence and weaker players are left defending against a significantly expanded array of threats from multiple antagonists.

AI-Driven Risks Across Key Industries

While AI is transforming security practices across industries, its implementation comes with sector-specific risks that require tailored solutions. Different industries face unique vulnerabilities that AI systems must address, making industry-specific risk management strategies essential.

Healthcare

In healthcare, AI is increasingly used for diagnostic support, operational optimization, and predictive analytics. However, the industry faces data privacy and patient safety challenges. AI systems that rely on sensitive patient data must comply with strict regulations like HIPAA, ensuring that data is anonymized and secure. Additionally, AI may be used to predict equipment failures or operational bottlenecks, but human oversight remains critical to handling complex, life-or-death decisions in clinical settings.

Finance

In finance, AI systems play a prominent role in fraud detection and risk assessment. By analyzing transaction data in real time, AI can detect anomalies and flag potentially fraudulent activities before they result in losses. However, these systems also introduce risks related to data integrity and algorithmic fairness. Biases in AI models could disproportionately impact certain customer groups, necessitating continuous monitoring and updates to avoid discrimination.

Manufacturing and Industrial

In manufacturing and supply chains, AI optimizes logistics, predicts machine failures, and ensures operational continuity. However, AI-driven automation increases the attack surface, with IoT and edge devices becoming more vulnerable to cyberattacks. In this sector, AI systems must predict internal risks and mitigate external disruptions, such as cyberattacks targeting global supply chains.

Tailoring AI risk management strategies to these industry-specific challenges is critical to harnessing AI's full potential while mitigating its risks.

2.3.4 The future of risk in the AI-driven world

The AI-driven world introduces a range of emerging risks that are both complex and multifaceted. From the vulnerabilities of IoT and edge devices to the challenges of global supply chains and potential adoption by non-state actors, enterprises must navigate a rapidly evolving threat landscape. While AI offers enormous potential for innovation and efficiency, it also creates new opportunities for disruption, especially as it intersects in the far future with other advanced technologies like quantum comput-

ing, which could render even the strongest encryption useless if advancements in encryption technology fail to keep pace (P. Mercer, personal communication, December 6, 2024). To address these risks, organizations must adopt a forward-thinking approach to security that anticipates and mitigates the consequences of AI's accelerating development. By understanding these emerging risks, enterprises can better position themselves to harness AI's benefits while safeguarding against its dangers.

2.4 AI's role in proactive risk management

In the modern enterprise landscape, the complexity of security threats continues to grow, driven by an increasing number of connected devices, evolving cyberattack methods, and globalized operations. Traditional reactive approaches to security are no longer sufficient in this fast-changing environment. To stay ahead of potential risks, enterprises will rely on artificial intelligence (AI) for more proactive risk management strategies. AI can support real-time threat detection, adapt security measures dynamically, and integrate predictive insights into decision-making and scenario planning.

2.4.1 Using AI for predictive threat detection and real-time monitoring

One of the most transformative applications of AI in proactive risk management is predictive threat detection. Unlike most current security approaches that primarily react to incidents after they occur, AI-based systems can analyze massive amounts of data from network activity, user behavior, and intelligence feeds to identify patterns indicative of emerging threats – whether human or machine-driven. This predictive capability will allow security teams to intervene before an incident escalates, minimizing damage and downtime. (Tribe AI, n.d.)

AI models leverage machine learning algorithms that learn from historical data and continuously improve over time, enhancing their ability to predict threats. This capability is especially useful for detecting subtle or slow-moving attacks that might evade traditional monitoring systems, such as advanced persistent threats (APTs) or consistent behavior patterns that build into potential threats over time. By identifying potential risks early, AI enhances real-time monitoring and allows security teams to mitigate threats proactively.

In addition to internal monitoring, AI plays a critical role in external threat intelligence by analyzing global threat data, including information from dark web forums, malware databases, social media, news, and public information sites. AI can quickly identify emerging threats or shifts in attacker tactics, techniques, and procedures, al-

lowing enterprises to adjust their strategies in real time. AI Risk Assessment specialist Paul Mercer notes the benefits of this AI assistance: “At a basic level, the ability to scrape data from trusted websites and conduct logical threat analysis can significantly save significant time and empower under-resourced security teams to communicate risk to business leadership” (P. Mercer, personal communication, December 6, 2024). In more advanced future cases, AI-driven systems will be able to dynamically modify security programs, enabling organizations to stay one step ahead of attackers. According to internal control specialist Adeleke Ayobami, “Automated systems can rapidly analyze and respond to security incidents, minimizing the window of vulnerability. From isolating compromised systems to applying necessary security patches, AI-driven incident response ensures a swift and coordinated reaction to emerging threats.” Even when automated agility is not appropriate, and human decision-making is needed, these insights can then be incorporated into a broader risk management strategy, enhancing an organization’s overall threat mitigation capabilities (Ayobami, 2024).

The scalability of AI-based security systems also allows organizations to respond effectively to increased demand during high-risk periods, such as major product launches, significant events, or public crises. AI-driven security platforms often integrate with a wide array of data sources. AI’s ability to process large volumes of data ensures that enterprises can continue to operate securely even as their attack surfaces expand. This integration enables a holistic approach to security that adapts not only to cyber threats but also to physical risks.

Table 2.3: Traditional vs AI-Augmented Risk Management Approaches.

Feature	Traditional Approach	AI-Augmented Approach
Threat Identification	Manual audits; alerts from known signatures	Continuous monitoring with anomaly detection
Response Time	Minutes to hours	Real-time or near-instant
Data Volume Handled	Limited	Petabyte-scale processing
Proactivity Level	Mostly reactive	Predictive and anticipatory
Role of Human Analysts	Central to all decisions	Augmented by AI insights; focused on edge cases

Beyond immediate threat detection and response, AI significantly impacts proactive decision-making and scenario planning in risk management. By analyzing both historical data and current trends, AI systems can generate predictive models that help enterprises anticipate a wide range of potential risks. These models allow organizations to simulate different scenarios and develop tailored strategies to mitigate the impact of those risks. For example, AI can predict the likelihood of specific disruptions—such

as cyberattacks, supply chain failures, or natural disasters—based on past incidents and emerging trends. AI can then provide recommendations on how to allocate resources, optimize disaster recovery plans, or adjust supply chain strategies to minimize potential losses. This type of scenario analysis helps organizations prepare for multiple risk scenarios, ensuring they are not caught off-guard by unforeseen events (Blessing et al. 2024)

Table 2.3 highlights the ways in which AI is revolutionizing proactive risk management by enabling enterprises to predict, detect, and respond to threats in real time while adapting to changing security landscapes. Its ability to integrate vast data sources, from internal systems to global threat intelligence, empowers organizations to make informed, proactive decisions and implement adaptive security measures. As AI-driven technologies continue to evolve, they will play an increasingly crucial role in helping enterprises navigate complex and dynamic risks, transforming risk management from a reactive process to a forward-looking strategy. Enterprises that leverage AI effectively will be in a better position to anticipate and mitigate threats, ensuring resilience in an increasingly unpredictable world.

Successful AI Implementations in Risk Management

Several industries have successfully integrated AI into their risk management frameworks, demonstrating the power of AI to improve security outcomes. In the energy sector, AI-driven systems are used to protect critical infrastructure such as power grids and oil refineries. By monitoring industrial control systems (ICS) and analyzing sensor data in operational technology (OT) environments, AI can detect anomalies that may indicate a cyberattack. Early detection of these anomalies allows operators to respond quickly and mitigate potential damage before it disrupts essential services.

Similarly, in healthcare, AI is being used to safeguard sensitive patient data. Hospitals leverage machine learning models to monitor access to electronic health records (EHRs) and detect unauthorized access attempts. These systems can flag suspicious behavior, such as accessing an unusually high volume of records outside regular working hours, and automatically initiate lockdown protocols to prevent data breaches.

In the retail and e-commerce sectors, AI tools are used to monitor supply chain security. By tracking shipments and identifying irregularities in inventory management, AI helps retailers detect potential fraud or theft before significant losses occur. AI systems can also predict supply chain disruptions by analyzing factors such as weather patterns, geopolitical risks, and vendor performance.

Finally, in the financial services industry, AI-driven systems have been deployed to detect fraud in real-time. By analyzing large volumes of transactional data, these systems identify patterns that may indicate fraudulent activity. As these machine learning models continue to evolve, their ability to adapt to new types of fraud becomes more sophisticated, allowing them to flag potential threats with greater accuracy.

2.5 Ethical and legal implications of AI in security

As AI systems take on more responsibility for decision-making in security, ethical considerations become more important. While AI brings numerous advantages in terms of efficiency and real-time threat detection, it also raises complex ethical and legal concerns that organizations must address. In an early in-depth look at AI ethics in 2020, Harvard Gazette writer Christina Pazzanese noted that "AI presents three major areas of ethical concern for society: privacy and surveillance, bias and discrimination, and perhaps the deepest, most difficult philosophical question of the era, the role of human judgment" (2020). As AI's role in security expands, enterprises need to ensure that its use is not only functional but also ethical and legally sound.

While AI can automate many aspects of security, it cannot make value-based decisions or account for the complex ethical implications of its actions. Human oversight is, therefore, essential in ensuring that AI-driven security systems align with ethical standards and organizational values. While specific to the auditing arena, Murikah, Nthenge, & Musyoka, in their paper *Bias and ethics of AI systems applied in auditing – A systematic review,* identify the issue that "AI cannot meet crucial requirements for ethical decision-making independently . . . [and] require careful oversight and collaboration" (2024). Expand that to the security arena and one can see that AI systems that monitor employees' behavior for security purposes could easily overstep boundaries and infringe on privacy rights if not properly governed. Enterprises must balance the introduction of AI into surveillance, facial recognition, and data analysis with respect for individual rights and compliance with privacy regulations. Human security professionals will play a crucial role in maintaining this balance by establishing ethical guidelines for AI systems and ensuring that AI is used responsibly.

One of the primary ethical concerns regarding AI in security is privacy. AI systems often require vast amounts of data to function effectively, particularly in predictive analytics and real-time monitoring. This data frequently includes sensitive personal information, from employee activities to customer behavior patterns. The aggregation of large data sets in AI systems makes them attractive targets for cybercriminals. Additionally, even without a threat of theft, the ability of AI systems to collect, analyze, and store personal data raises concerns about how the information is stored and protected, and also how it could be leveraged for unauthorized uses. raises significant privacy issues. As noted by the National Institute of Standards and Technology (NIST), "The rapid proliferation of Artificial Intelligence (AI) promises significant value for industry, consumers, and broader society, but as with many technologies, new risks from these advancements in AI must be managed to realize its full potential" (Megas, 2024). AI Risk Assessment specialist Paul Mercer notes that "Anonymization of data is key to establishing and maintaining trust in these systems" (P. Mercer, personal communication, December 6, 2024).

Further, AI-driven surveillance systems, such as facial recognition, can infringe on individual privacy if not adequately managed. These systems can track movements

and behaviors, sometimes without the explicit consent of the individuals under surveillance. Additionally, AI-powered devices and services, such as virtual assistants and IoT products, continuously collect personal information, raising concerns about intrusive data gathering and unauthorized access by third parties. The Brookings Institution highlights that "Artificial intelligence likely will accelerate this trend. Much of the most privacy-sensitive data analysis today–such as search algorithms, recommendation engines, and adtech networks–are driven by machine learning and decisions by algorithms. As artificial intelligence evolves, it magnifies the ability to use personal information in ways that can intrude on privacy interests by raising analysis of personal information to new levels of power and speed" (West, 2018).

Another major ethical issue in AI security is the risk of bias in AI algorithms. AI systems are trained on historical data, which means they can inadvertently learn and reinforce the biases present in that data. This issue is particularly concerning in security contexts where AI-driven decisions (such as identifying potential threats or allocating resources) can have significant real-world consequences. A good example of the possible ethical impact of AI bias is in facial recognition systems. These systems often perform less accurately for some demographic groups, particularly women and people of color. If security or law enforcement deploys facial recognition to identify suspects in incidents, biased algorithms could result in discriminatory outcomes, such as higher rates of false positives for minority groups (Sutaria, 2022).

While AI offers transformative potential for security systems, enterprises must navigate complex ethical, legal, and social challenges. Across the globe, there is an increasing emphasis on data protection laws and AI regulations, such as the European Union's General Data Protection Regulation (GDPR) and the proposed Artificial Intelligence Act. These laws aim to regulate how personal data is collected, processed, and used by AI systems, and non-compliance can result in severe penalties. As we move into a more AI-enabled future, human security professionals will play a critical role in addressing these concerns by establishing ethical guidelines, ensuring compliance, and maintaining the balance between technological innovation and respect for individual rights.

2.5.1 The need for strong AI governance and transparency

As organizations increasingly rely on AI for security operations, ensuring AI governance is critical to managing the ethical, legal, and operational challenges that accompany this technology. Examples of the applicable risks and challenges are outlined in table 2.4. Effective AI governance frameworks provide a roadmap for transparent and accountable decision-making, ensuring that AI systems operate ethically, transparently, and in alignment with organizational values. For instance, the NIST AI Risk Management Framework (AI RMF) in the United States provides voluntary guidelines to help organizations manage AI risks and emphasizes trustworthiness and responsible development across sectors

(NIST, n.d.). In contrast, the European Union's AI Act, which came into force in August 2024, mandates a risk-based regulatory approach, categorizing AI systems by their potential impact and imposing corresponding obligations (European Commission, 2024). While their differences highlight how regional priorities influence AI governance, with the U.S. adopting decentralized strategies and the EU favoring comprehensive legislation (Li, 2024), their focus is clearly on a governance approach. As pointed out by Allen and Bapst in their 2024 book on Cyber Risk Management, "Governments worldwide clearly recognize the urgent need for a formal approach to the new and unpredictable risks Al represents" (Allen & Bapst, 2024). Global governments are fully engaged in this issue and private enterprises will need to follow suit to appropriately manage AI-related risk.

For private organizations who want to adopt similar approaches to AI risk, governance frameworks will need to include clear protocols for data management, ensuring that the data AI systems rely on is both ethically sourced and securely managed. Paul Mercer, an early business adopter of AI in the Risk Assessment space mentions "Our company recognizes the importance of explaining, in non-technical terms, how the LLM assesses threats and the data it utilizes to do so. This transparency is essential for building trust with stakeholders" (P. Mercer, personal communication, December 6, 2024). Governance will also require the implementation of continuous audits of AI systems to detect any biases or unintended behaviors, especially in areas like threat detection and surveillance. Transparency is essential here: enterprises must be able to explain AI decisions not only to internal stakeholders but also to regulators and the public. True transparency includes maintaining explainable AI principles and ensuring that AI-generated decisions can be understood and evaluated by human operators (Sheikh Hassani, 2023).

Table 2.4: Categories of AI Risk and Suggested Governance Approaches.

Risk Category	Description	Suggested Controls
Data Bias	Inherited from training data; may produce unfair results	Bias audits; diverse datasets; fairness metrics
Model Drift	Model performance degrades over time	Continuous monitoring; retraining policies
Explainability	AI decisions are opaque	Use of interpretable models; explainability tools
Misuse	AI used for malicious purposes	Access controls; red-teaming; ethical guardrails
Regulatory Compliance	Non-compliance with AI laws (EU AI Act, etc.)	Policy alignment; legal review; AI ethics boards

Governance also plays a crucial role in mitigating AI bias. AI models are trained on historical data. Therefore, they can inadvertently reinforce existing biases. A strong governance structure includes regular checks and updates to ensure that AI systems evolve with ethical considerations in mind, addressing concerns about fairness, privacy, and accountability (Allen & Bapst, 2024). In the future, organizations will need to collaborate across all business areas to set standards for AI governance, ensuring that AI systems operate both fairly and consistently. This governance should be flexible enough to adapt to new threats and regulatory changes, maintaining the balance between AI innovation and human oversight.

As AI continues to reshape enterprise security, security leaders must carefully manage the ethical and legal implications of its use with input from all experts inside and outside the enterprise who are familiar with the ethical and legal implications of adopting AI and automated technologies. Privacy concerns, algorithmic bias, and regulatory compliance are critical challenges that organizations must address to ensure the responsible deployment of AI-driven security tools (Rooney, 2024). By prioritizing ethical governance and oversight, protecting personal data, and adhering to emerging regulatory standards, enterprises can leverage AI's benefits while minimizing its risks.

2.6 Human-AI collaboration: The new security paradigm

The integration of artificial intelligence into security operations has introduced profound changes in how organizations protect themselves from evolving threats. In this new era, the relationship between human expertise and AI-driven automation will be not one of competition but collaboration. While AI offers unmatched efficiency, speed, and data processing capabilities, it lacks the contextual understanding, ethical judgment, and creative problem-solving that humans bring to security decision-making. As organizations continue to deploy AI in security, the balance between automation and human intuition becomes a critical factor in ensuring robust and adaptive security postures.

2.6.1 Automation vs. Human intuition: Balancing AI-driven automation with human decision-making in security

AI excels at automating repetitive and data-intensive tasks, from monitoring network traffic for anomalies to scanning vast amounts of video footage for suspicious activity. These automated processes free human security professionals from time-consuming, mundane tasks, allowing them to focus on higher-level strategic decision-making. AI-

powered systems can process vast datasets far more quickly than any human team, identifying patterns that would otherwise go unnoticed. For instance, AI can detect and respond to network intrusions in real time, flagging suspicious behavior and initiating countermeasures before human operators are even aware of the breach.

However, while automation increases efficiency, there are limits to what AI can achieve on its own. Technology and security editor Kirsten Doyle notes that "While AI systems excel at analyzing data and providing insights, human knowledge and expertise are needed to make strategic decisions based on this information. People are essential for prioritizing responses, weighing up risks, and deciding how best to allocate resources in the face of a threat" (2024). Additionally, AI systems operate based on predefined algorithms and learn from historical data, which means they are only as effective as the data used to train them. In contrast, human intuition allows for adaptability in situations where predefined rules do not apply or where unprecedented threats emerge. A human operator, for example, might notice subtle contextual details or anomalies that an AI would overlook, especially in scenarios where creative problem-solving is required. Balancing the speed and precision of AI with the critical thinking and ethical judgment of human operators creates a synergistic approach to security (Doyle, 2024).

2.6.2 The future role of security professionals

As AI continues to transform the security landscape, the role of security professionals will change. They will need to adapt to a more proactive role and leverage AI to anticipate threats, manage risks, and optimize security operations. Allen and Bapst identify that "Security has a unique and critical role to play in addressing the rapidly evolving attack surfaces and emerging risks presented by Al and other digital technologies" (2024).

The future security professional will require more than just technical security expertise; they will need more robust business acumen and a deep understanding of interdisciplinary fields. Security strategies must align with broader business objectives. As the security landscape evolves, it is crucial to recognize that "security professionals play a critical and leading role in ensuring security considerations are integrated into the overall business strategy to achieve desired effects" (Jayasundera, 2024). This alignment ensures that as AI-driven security tools become more integrated into enterprise operations, security professionals can effectively collaborate with departments such as legal, compliance, and operations to ensure that security measures support the organization's goals.

Furthermore, risk management is becoming an increasingly important skill. AI offers powerful tools for identifying and mitigating risks. However, it is up to human security professionals to interpret AI-generated insights and make informed decisions based on an understanding of the broader business context.

Human-Machine Collaboration: AI and Human Security Professionals Work in Tandem
The success of AI in security often hinges on the collaboration between human experts and AI systems. Numerous case studies highlight how this collaboration enhances security outcomes, particularly in complex, high-risk environments where neither AI nor humans can effectively operate alone.

Hospitality
In hospitality, AI is used to automate routine operations such as booking management and guest inquiries, freeing staff to focus on providing personalized service. For example, hotels use AI-driven concierge services like IBM Watson to answer guest questions, make recommendations, and manage bookings, while human staff handle more nuanced requests or problem-solving. AI also assists in analyzing guest data to create tailored experiences, such as personalized itineraries, based on historical preferences. During large-scale events or peak seasons, AI systems optimize room rates and staff deployment, but human managers still guide strategic decisions around guest experience and loyalty programs. In industries such as hospitality, the "human touch" is critical to guest experience, and AI and automation support the staff and enable them to spend less time on mundane tasks and more time face-to-face with customers,

Healthcare
AI-human collaboration is crucial in healthcare, particularly in diagnostic and operational efficiency. AI-powered imaging tools assist radiologists by analyzing scans for subtle patterns, helping detect conditions like lung nodules or neurological changes more quickly and accurately than humans alone. However, radiologists remain in control of final diagnoses and treatments, using AI as a supportive tool rather than a replacement. In hospital operations, AI predicts patient volumes and optimizes resource allocation (e.g., staffing and equipment), but human oversight ensures that patient care decisions are sensitive to unique clinical contexts. Healthcare remains a person-to-person arena but one that can benefit significantly from the pattern recognition and data recall provided by automated systems.

Energy and Utilities
In the **energy sector**, AI-driven systems monitor critical infrastructure, such as power grids, for signs of cyberattacks or equipment failure. These systems can detect anomalies in operational technology (OT) environments—like sudden voltage changes—that may signal an impending failure or cyber intrusion. Human operators then intervene to assess the situation and initiate appropriate actions, such as rerouting power or deploying repair crews. In these high-stakes environments, AI's ability to provide early warnings allows human operators to make more informed decisions, preventing potential blackouts or other crises.

2.6.3 Preparing security teams for AI integration

As AI becomes more embedded in security operations, organizations must invest in training and development programs that prepare security teams for AI integration. The rapid evolution of AI technologies requires continuous learning, as security professionals must not only understand how to use AI tools but also how to interpret their outputs and make informed decisions based on AI-generated insights. This is beginning to be seen as not only "good practice," but also is being codified in regulatory requirements. In 2024, the New York State Department of Financial Services issued guidance emphasizing the importance of updating risk assessments and implement-

ing incident response plans to address AI-related threats. One example of the types of training requirements included is:

> *The organization's cybersecurity personnel and relevant TPSP personnel must undergo specific cybersecurity training, which delves into the threats of AI-generated cyberattacks, how AI can be used to improve cybersecurity, how to secure and defend AI systems for cyberattacks and how to design and develop AI systems securely;*

This early regulatory salvo reinforces the need for security teams to understand the ethical and regulatory implications of AI in security (Eng & Kuruvilla, 2024).

Security professionals will need to develop hybrid skills that allow them to manage AI technologies alongside traditional security roles. Ongoing education and training will be critical to ensuring that security professionals are equipped with the skills necessary to manage AI-driven systems effectively.

Organizations should prioritize comprehensive training programs that encompass both the technical aspects of AI—such as machine learning and data analysis—and the ethical and regulatory implications of AI in security. Practical, hands-on experience with AI systems is essential for security teams to effectively manage real-world AI deployments. This includes training on monitoring AI systems for signs of malfunction, interpreting AI-generated data, and adjusting AI models based on emerging threats.

Already the market is preparing for this need. The SANS Institute offers courses like "SEC595: Applied Data Science and AI/Machine Learning for Cybersecurity Professionals," which focus on the practical application of AI in cybersecurity (SANS Institute, n.d.). Additionally, the Certified AI Security Professional (CAISP) course offers an in-depth exploration of AI supply chain risks, equipping professionals with the knowledge and skills to identify, assess, and mitigate these risks (Practical DevSecOps, n.d.). These programs are designed to equip security professionals with the necessary skills to manage AI-driven systems effectively and show that the industry is already anticipating this need.

By investing in such training initiatives, organizations can ensure that their security teams are well-prepared to handle the complexities of AI deployments, thereby enhancing their overall security posture.

The collaboration between humans and AI represents a new security paradigm, one that balances the speed and precision of automation with the intuition, ethics, and contextual understanding that only human operators can provide. As AI continues to evolve, the role of security professionals will shift, requiring more substantial business acumen, interdisciplinary expertise, and hybrid skills that integrate AI technologies with traditional security practices. Organizations must invest in continuous learning and development to prepare security teams for this transition, ensuring that human-AI collaboration leads to vigorous, more adaptive security strategies.

2.7 Conclusion

As artificial intelligence continues to evolve, its dual role as both a tool and a threat in enterprise security is becoming increasingly evident. On the one hand, AI offers unprecedented capabilities for predictive threat detection, real-time monitoring, and adaptive responses to emerging risks. Enterprises that leverage AI's analytical power can stay one step ahead of cyberattacks and enhance their physical security measures. AI systems can process vast amounts of data at speeds far beyond human capabilities, uncovering patterns and anomalies that allow for faster, more informed decision-making.

However, this power comes with risks. AI is also a significant threat vector, capable of being exploited by malicious actors to automate and scale attacks, create sophisticated phishing schemes, or generate deepfakes and misinformation campaigns. Vulnerabilities in AI systems, particularly when integrated with IoT and edge devices, expand the attack surface for enterprises, making it essential to approach AI deployment with caution and vigilance.

Despite AI's impressive capabilities, human security professionals remain indispensable in ensuring that AI-driven systems function safely, ethically, and effectively. AI systems can automate many tasks, but they lack the nuanced judgment, ethical consideration, and creative problem-solving that human operators bring to complex security environments. In scenarios where AI systems flag potential threats, it is still the responsibility of human professionals to verify those findings, manage responses, and make strategic decisions that align with business goals.

Moreover, as AI systems grow more sophisticated, the role of human oversight becomes even more critical. Without careful governance, AI algorithms can introduce bias, violate privacy, or fail to adapt to novel threats. To mitigate these risks, organizations must establish robust governance frameworks that ensure human control over AI's outputs. These frameworks should include regular audits of AI decision-making processes, transparency around how AI models are developed and trained, and clear accountability structures to ensure that human operators retain ultimate authority.

In the future, the successful integration of AI with human expertise will influence and define security risk management. Enterprises that prioritize AI governance—ensuring that AI systems are transparent, accountable, and aligned with organizational ethics—will be better positioned to navigate the complex landscape of AI-driven risks and opportunities.

As AI adoption increases, organizations will need to continuously adapt their security practices to manage the technological and ethical challenges associated with it. The path forward includes investing in ongoing training for security professionals, developing hybrid teams that blend AI skills with traditional security expertise, and fostering a culture of collaboration between humans and machines. By doing so, enterprises can ensure that they not only stay ahead of AI-driven threats but also fully capitalize on AI's opportunities to enhance their security operations.

AI's future role in enterprise security is both powerful and complex. While it offers unmatched potential for improving risk management and threat detection, it also introduces new risks that demand careful management. The future of enterprise security will depend on maintaining a careful balance between leveraging AI's strengths and ensuring that human professionals remain at the helm, guiding AI's deployment with ethical oversight and strategic vision.

References

Allen, B., Bapst, B., & Hicks, T. A. (2024). *Building a cyber risk management program: Evolving security for the digital age*. O'Reilly Media.

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Al Yami, M. S., Al Harbi, S., & Albekairy, A. M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC medical education, 23(1), 689. https://doi.org/10.1186/s12909-023-04698-z

Andrews, C. H., Cohen, J. M., Lew, D., Stark, T., Diouf, M., & Mano, S. (2024, March 14). *DOJ doubles down on warnings against AI misuse*. White & Case LLP. https://www.whitecase.com/insight-alert/doj-doubles-down-warnings-against-ai-misuse

Ayobami, A. (2024, June 28). *Leveraging AI for information and cybersecurity*. ISACA. https://www.isaca.org/resources/news-and-trends/industry-news/2024/leveraging-ai-for-information-and-cybersecurity

Beato, F., Rohland, L., & Van Gogh, M. (2024, June 20). *Why transport and supply chain ecosystems need to be cyber secured*. World Economic Forum. https://www.weforum.org/stories/2024/06/transport-supply-chain-ecosystems-cyber-resilience/

Black, J. (2024, September 9). The potential of generative artificial intelligence to revolutionize physical security. *Security Industry Association*. https://www.securityindustry.org/2024/09/09/the-potential-of-generative-artificial-intelligence-to-revolutionize-physical-security/

Blessing, M., Kolawole, W., & Owen, J. (2024). *The impact of AI-powered threat detection systems on modern cybersecurity practices*.

Control Risks. (2024, October 30). *How artificial intelligence is lowering the barrier to cybercrime*. Control Risks. https://www.controlrisks.com/our-thinking/insights/how-artificial-intelligence-is-lowering-the-barrier-to-cybercrime

De Angelo, D. (2024, May 15). *The dark side of AI in cybersecurity — AI-generated malware*. Palo Alto Networks. https://www.paloaltonetworks.com/blog/2024/05/ai-generated-malware/

Doyle, K. (2024, October 17). Stronger together: AI and human collaboration in the battle against evolving threats. *Information Security Buzz*. Retrieved November 8, 2024, from https://informationsecuritybuzz.com/ai-and-human-collaboration-in-battle/

Eng, D. M., & Kuruvilla, S. (2024, November 15). New York Department of Financial Services provides AI cybersecurity guidance: What it means. *Reuters*. Retrieved from https://www.reuters.com/legal/legalindustry/new-york-department-financial-services-provides-ai-cybersecurity-guidance-what-2024-11-15/

European Commission. (2024). *AI Act enters into force*. Retrieved from https://commission.europa.eu

Forbes Technology Council. (2023, May 18). *Mitigating dark web risks: The role of AI and machine learning*. Forbes. Retrieved from https://www.forbes.com/councils/forbestechcouncil/2023/05/18/mitigating-dark-web-risks-the-role-of-ai-and-machine-learning/

Fortinet. (n.d. -a). Artificial intelligence in cybersecurity. *Fortinet*. Retrieved November 22, 2024, from https://www.fortinet.com/resources/cyberglossary/artificial-intelligence-in-cybersecurity

Fortinet. (n.d. -b). *Top IoT device vulnerabilities: How to secure IoT devices*. https://www.fortinet.com/resources/cyberglossary/iot-device-vulnerabilities

Heiding, F., Schneier, B., & Vishwanath, A. (2024, May 30). AI will increase the quantity and quality of phishing scams. *Harvard Business Review*. https://hbr.org/2024/05/ai-will-increase-the-quantity-and-quality-of-phishing-scams

Jayasundera, S. (2024, January 5). *Ways to enhance strategic influence of security professionals within organization*. Disaster Recovery Journal. Retrieved from https://drj.com/journal_main/ways-to-enhance-strategic-influence-of-security-professionals-within-organization/

Li, L. (2024, October 21). *Comparing EU and US AI legislation: Déjà vu to 2020*. Reuters. Retrieved from https://www.reuters.com/legal/legalindustry/comparing-eu-us-ai-legislation-dj-vu-2020-2024-10-21/

Megas, K. (2024, September 19). *Managing cybersecurity and privacy risks in the age of artificial intelligence: Launching a new program at NIST*. National Institute of Standards and Technology. https://www.nist.gov/blogs/cybersecurity-insights/managing-cybersecurity-and-privacy-risks-age-artificial-intelligence

Murikah, W., Nthenge, J. K., & Musyoka, F. M. (2024). Bias and ethics of AI systems applied in auditing: A systematic review. *Scientific African*, *25*, e02281. https://doi.org/10.1016/j.sciaf.2024.e02281

National Security Agency, Federal Bureau of Investigation, & Cybersecurity and Infrastructure Security Agency. (2023). *Contextualizing deepfake threats to organizations* (Cybersecurity Information Sheet). U.S. Department of Defense. https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF

NIST. (n.d.). *AI Risk Management Framework (AI RMF)*. Retrieved from https://www.nist.gov

OWASP Foundation. (2024). *OWASP Top 10 for LLM & Generative AI Security*. OWASP. https://genai.owasp.org/llm-top-10/

Pazzanese, C. (2020, October 26). *Ethical concerns mount as AI takes bigger decision-making role*. Harvard Gazette. https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/

Polemi, N., Praça, I., Kioskli, K., & Bécue, A. (2024). Challenges and efforts in managing AI trustworthiness risks: A state of knowledge. *Frontiers in Big Data*, *7*, Article 1381163. https://doi.org/10.3389/fdata.2024.1381163

Practical DevSecOps. (n.d.). Certified AI Security Professional (CAISP). Retrieved from https://www.practical-devsecops.com/certified-ai-security-professional/

Rooney, P. (2024, November 4). *CIOs look to sharpen AI governance despite uncertainties*. CIO. https://www.cio.com/article/3595801/cios-look-to-sharpen-ai-governance-despite-uncertainties.html

Sadhu, P. K., Yanambaka, V. P., & Abdelgawad, A. (2022). Internet of Things: Security and solutions survey. *Sensors*, *22*(19), 7433. https://doi.org/10.3390/s22197433

SANS Institute. (n.d.). SEC595: Applied data science and AI/machine learning for cybersecurity professionals. Retrieved from https://www.sans.org/ai/

Sheikh Hassani, H. (2023, September 14). *AI governance: Frameworks, tools, best practices*. DataCamp. https://www.datacamp.com/blog/ai-governance

Schneier, B., & Raghavan, B. (2024). Seeing like a data structure. Harvard Kennedy School, Belfer Center. https://www.schneier.com/essays/archives/2024/05/seeing-like-a-data-structure.html

Sutaria, N. (2022, August 29). *Bias and ethical concerns in machine learning*. ISACA Journal, 4. https://www.isaca.org/resources/isaca-journal/issues/2022/volume-4/bias-and-ethical-concerns-in-machine-learning

Tariq, U., Ahmed, I., Bashir, A. K., & Shaukat, K. (2023). A critical cybersecurity analysis and future research directions for the Internet of Things: A comprehensive review. *Sensors*, *23*(8), 4117. https://doi.org/10.3390/s23084117

Tribe AI. (n.d.). *AI for cybersecurity: How online safety is enhanced by artificial intelligence*. Tribe AI. https://www.tribe.ai/applied-ai/ai-for-cybersecurity

Tziakouris, G. (2023, August 14). *The rise of AI-powered criminals: Identifying threats and opportunities*. Cisco Talos Blog. https://blog.talosintelligence.com/the-rise-of-ai-powered-criminals/

West, D. M. (2018, February 12). *Protecting privacy in an AI-driven world*. Brookings Institution. https://www.brookings.edu/articles/protecting-privacy-in-an-ai-driven-world/

World Economic Forum. (2024). *The global risks report 2024* (19th ed.). World Economic Forum. https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf

Zacharia, Z. G., & Center for Supply Chain Research at Lehigh University. (2024). *Lehigh Business Supply Chain Risk Management Index: 4th Quarter Report*. Council of Supply Chain Management Professionals. https://business.lehigh.edu/sites/default/files/2024-10/LRMIReport20244thQuarterReportv2.pdf.

Matthew LaGana

3 Securing the future: Managing cybersecurity risks in the age of artificial intelligence

3.1 Introduction

AI is transforming cybersecurity. It enhances defense but also introduces new threats. When used in business operations, AI's positive and negative aspects as a system control and weapon present unique risks to systems and overall data integrity.

The application of AI in cybersecurity has been previously seen as a way of enhancing detection and automating response to threats, which would increase the speed of detecting threats. AI can analyze patterns and detect anomalies within a short time, a task that is impossible for human analysts. Machine learning detects threats quickly, alerting security teams and limiting impact. Nevertheless, the capabilities that AI has reinforced in the field of cybersecurity are also being used by adversaries. Cybercriminals are using AI to automate cyberattacks and make them more efficient, effective, and fast (FBI, 2024). These new and sophisticated AI-assisted attacks are a major challenge to traditional cybersecurity measures and defense practices. This chapter aims to discuss the emerging risks of AI in cybersecurity, present precautions that can be taken, and explain how organizations can develop their AI strategy to enhance overall cyber resilience.

AI advancements have expanded the cybersecurity threat landscape. With that, the integration of AI into our cybersecurity postures offers numerous benefits. For instance, the automation of threat detection and real-time response. AI can process enormous amounts of data in a short time, which is not possible with human analysts, which leads to more effectively mitigating risks or threats. This has proved to be useful in network security, where AI-based systems are used to actively analyze traffic and prevent intrusions. But this same power that helps in our defense also helps the cybercriminals. Adversaries are employing AI to design sophisticated schemes to launch attacks, which is a challenge to cybersecurity professionals. The threat has shifted from simple attacks to dynamically coded malicious software to complex and evolving AI-inspired threats (Gartner, 2023).

3.2 AI powered cyber attacks: A new threat

AI-driven threats have evolved from simple attacks to complex, adaptive ones. As AI develops, so do the ways that adversaries can exploit it. The best example is adversaries using AI to design better and more targeted phishing emails.

https://doi.org/10.1515/9783111595306-003

The following are the different types of attacks that can be launched using AI:

AI-Powered Phishing Campaigns: A growing threat, phishing campaigns have become very sophisticated due to the fact that AI can collect and analyze large amounts of data. Before, phishing attempts were done through generic messages, while AI phishing uses social media profiles, email histories, and personal communication patterns to come up with very specific and personalized messages. These attacks are more realistic and almost impossible to differentiate from real ones. The phishing campaigns that are powered by AI can focus on the trust and knowledge of the organizational hierarchy. In the case of spear phishing, attackers send emails that appear to be from other employees of the organization or high-ranking executives. These AI-created emails can mimic an individual's language and writing style to ensure that the recipient will click on the link or download the infected attachment. To prevent such kinds of attacks, companies must update their threat detection systems to recognize not only traditional phishing attempts but also the slight behavioral and linguistic features of AI-generated attacks. This will involve the use of AI detection systems and human analysis to ensure that organizations can identify new threats before they cause a data breach or other related cyber events (TechTarget, 2024).

AI-driven malware has also increased in complexity. One of the most alarming uses of AI in cyberattacks is the adaptive malware that can learn its way through an organization's systems. Malware that is controlled by AI can learn the organization's defenses and change its actions to avoid detection. Unlike traditional malware that uses a fixed approach in attacking by using certain patterns or methods, AI malware can change its code to avoid detection by signature-based methods and other standard cybersecurity tools.

Recent studies (CrowdStrike, 2023) performed by prominent technology companies have delved into how AI helps attackers develop their malware. For instance, Machine Learning algorithms could be used to update the malware's approaches according to its effectiveness or failure, making it more active and persistent in avoiding detection. This poses a major challenge to cybersecurity professionals because the methods that were previously used to detect malware, such as through signatures, are no longer effective. New approaches like behavioral analysis and anomaly detection-based AI must be adopted to deal with these new and improved threats.

3.3 Malicious GPTs

A generative AI model can produce intelligent text in response to user prompts. This technology, however, can also be altered by attackers (referred to as malicious GPT) when an altered version of the GPT produces harmful or deliberately misinformed outputs.

In the context of cyberattacks, a malicious GPT can generate attack vectors, such as malware, or supporting attack materials, such as fraudulent emails or fake online content, to advance an attack.

3.4 Ransomware attacks

AI-enabled ransomware is a type of ransomware that leverages AI to improve its performance or automates certain aspects of the attack path. For example, AI can be leveraged to research targets, identify system vulnerabilities, or encrypt data. AI can also be used to adapt and modify the ransomware files over time, making them more difficult to detect with cybersecurity tools.

3.5 AI for attacking biometric security systems

AI is also being used to subvert strong authentication factors like biometric identification. These are systems that are being implemented to use fingerprints, facial recognition, and other biometric characteristics to lock devices and networks and are vulnerable to AI-assisted attacks like deepfakes. Deepfake is a technology that uses machine learning to create realistic audio and video fakes that can fool biometric authentication systems and gain access to secured areas. Deepfakes can trick facial recognition into granting unauthorized access. In some cases, the deepfake audio can resemble the CEO's voice and thereby trick systems (and employees) into releasing access to important data or money. As the use of biometric systems increases, it is crucial to ensure that the systems are secured against AI-assisted attacks. This could entail employing other forms of authentication, like dual-factor authentication, or using other security measures in conjunction with biometric data, such as behavioral biometrics, to protect against deepfake attacks (CrowdStrike, 2025).

3.6 Employee awareness training

Organizations cannot underestimate the value and power of a strong cybersecurity awareness and education program. By incorporating AI-based attacks into the existing security training curriculum, users can learn how to spot and defend against AI-powered attacks. By using simulated attacks that focus on how realistic and convincing AI-enabled attack techniques can be, users are safely exposed to the social engineering techniques cybercriminals are actively using, such as deepfake chat and audio-based attacks.

To protect against adversarial AI attacks, we need to train our teams to recognize suspicious activity or outputs related to AI or Machine Learning-based systems. In this realm, the training systems can also identify individuals who are more prone to certain types of attacks and automatically align training to address these specific attacks. By identifying those individuals proactively, the probability of a successful social engineering attack is further reduced. With that, organizations can also use AI-based training systems to maximize the impact and effectiveness of training.

3.7 Data privacy and integrity risks in AI systems

The large amounts of data used and created by AI systems are a double-edged sword. On the one hand, there is no way AI can work without data; on the other hand, these datasets are a major threat to data privacy and integrity. If not well managed, AI systems can become a means of data theft and a source of data manipulation attacks.

3.8 Data poisoning: AI risk

Influencing the way AI models think, a major risk in AI-based systems is data poisoning. This attack occurs when the adversary injects malicious data into the training set of an AI model, such that the model generates incorrect or inappropriate output when deployed. For instance, an adversary may introduce some incorrect data into the training set and thus influence the AI model into making the wrong decision. In healthcare, this could mean wrong diagnoses; in autonomous vehicles, it could mean fatal errors in navigation. Since AI systems are primarily data-driven to a large extent, organizations must have strong controls on the data used for training. This entails the use of strong validation procedures to detect any attempt at data manipulation and only allowing data that has been collected and verified as accurate to be used for training models.

3.9 Bias in AI

The ethical issues and challenges of AI systems stem from their lack of inherent neutrality. They learn from data, and if the data they are trained on is biased, the model will also be biased. This can result in unequal treatment of individuals and can be particularly detrimental in critical areas such as employment, law enforcement, and credit. For instance, it has been established that AI-based hiring algorithms prefer male candidates over their female counterparts because of the historical data used.

To ensure that organizations do not perpetrate inequality through their AI systems, it is crucial to carry out regular bias audits on the AI systems. This entails making sure that the data used to train the AI models is diverse, inclusive, and has no bias. AI systems should also be explainable so that the user can understand how a decision has been made and challenge undesirable or unfair outcomes.

3.10 Adversarial AI-based attacks

A weakness of AI systems, adversarial machine learning attacks, is yet another major challenge to AI-based platforms. These attacks are based on modifying input data in such a way that an AI system would give the wrong output or decision. For example, a small change to an image that looks harmless may cause a facial recognition application to misidentify a person. Another example is adversarial attacks on autonomous vehicles, which are a major concern in particular. For instance, early in the journey of autonomous vehicles, it was revealed that small stickers could be attached to road signs to confuse the AI built into the self-driving mode of these cars. This meant that the car would interpret a stop sign as a yield sign, which could be fatal in real-world traffic. Other AI-driven vehicles and other intelligent systems are also prone to such attacks. Current research is on the development of AI systems that can identify and counter adversarial patterns, but as AI advances, the attacks will become more sophisticated as well. This means that the security of AI systems is not only a one-off job but rather requires constant assessment and enhancement.

3.11 Industry-specific case studies

AI risks are being identified across various sectors, as they all come with distinct cybersecurity risks. Let's now see how AI is being applied in various fields and what risks are involved in it.

Healthcare: Some of the opportunities and challenges of using AI in treating diseases and protecting patients' information. The use of AI in healthcare is transforming the way diseases are diagnosed and treated, but it also implies severe risks for patients' data. Just think what would happen if attackers identified and exploited vulnerabilities in an AI-powered system used by a healthcare provider to analyze patient information. If an automated diagnostic system were to be manipulated, it could expose vital patient records and misdiagnose critical conditions.

Financial Services: AI for fighting fraud and managing risk. Fraud detection is another area where AI is widely used in the financial services industry. Nevertheless, the application of AI may lead to biases and errors in the algorithm. An AI-based

fraud detection system could incorrectly flag legitimate transactions, causing disruptions for customers. and cause a bank or financial institution to block funds. This, therefore, demonstrates the need for having human involvement in the decision-making process of AI models in order to avoid negative effects.

3.12 The impact of artificial intelligence on strengthened security measures

A common use of AI for threat identification lies in intrusion detection systems (IDS), which are designed to pick up on abnormalities by learning from security breaches and adapting to new attack strategies. For instance, anomaly detection systems powered by artificial intelligence detect patterns in how employees log into their accounts. As a result, an organization can uncover potential advanced persistent threats (APT) by alerting on unusual use of access that differs from an established baseline. An APT can be disastrous for an organization, as it can avoid or evade detection for several months. Furthermore, AI strengthens security in connected devices by identifying threats and vulnerabilities through correlation engines. These AI-powered engines can take disparate log events that are seemingly harmless in and of themselves and translate them into meaningful data that correlates to known attacks or attack behavior.

3.13 Automated response to incidents

The incorporation of AI into incident response systems plays a role in speeding up response times and limiting the harm caused by cyberattacks. AI-driven systems can autonomously quarantine infected devices and address security flaws by revoking compromised credentials or applying patches without intervention. This quick responsiveness becomes critical in minimizing the repercussions of incidents and data breaches, as well as other forms of system breaches. During an attack situation, automated containment can be vital in stopping an attack in its tracks, as AI systems are capable of identifying malicious actions and isolating the impacted systems from the network to halt any further spread of the attack while also notifying security teams promptly. This quick reaction allows human teams time to investigate and respond to the cyberattack effectively.

3.14 A balancing act

Although automated incident response proves to be quite efficient, there are instances where it may incorrectly identify actions as threats (known as false positives). For example, routine software updates or maintenance activities could be mistakenly tagged as bad behavior. This can be incredibly disruptive to an organization and cause system outages and impact the user community. To address this issue, AI systems need to be adjusted and paired with learning processes to evolve and reduce mistakes over time.

3.15 Enhancing security through predictive analysis

In the realm of threat intelligence are cybersecurity advancements in data analytics using AI technology. This involves Machine Learning algorithms, studying past data to anticipate upcoming security risks based on identified patterns from previous cyber incidents. The ability to anticipate attacks helps organizations be proactive in their defense strategies by allocating resources effectively and implementing preventive measures ahead of time. Let's use the following example to illustrate this point. A cybersecurity company uses analytics driven by AI to predict an increase in phishing attempts around tax season. This foresight enables companies to enact tailored security measures and awareness initiatives. The AI model studies attack patterns to pinpoint targets and methods of attack.

3.16 Strategies to reduce cybersecurity risks posed by AI technology

Organizations must develop clear frameworks and strategies to effectively manage cybersecurity risks arising from advancements in AI technology. A key approach is adopting a risk-focused cybersecurity strategy. This involves identifying critical assets, assessing potential threats, and implementing defense measures tailored to the risk levels associated with each AI system component. For example, a company using an AI-driven customer data analysis platform should prioritize securing this system by allocating appropriate resources and implementing targeted protections.

The framework also needs to outline duties for handling cybersecurity threats linked to AI. Important components of AI governance encompass data management. Organizations need to make sure that their AI systems follow the rules laid out in regulations, like GDPR or the AI Act, by checking for security and compliance to steer clear of fines and legal issues. Although AI technology offers capabilities, it is essential

to remember that human supervision should not be replaced entirely. One of the advantages of AI in the field of cybersecurity is its capacity to enhance judgment rather than supplant it completely. Cybersecurity teams should collaborate with AI systems in verifying discoveries and guarantee that automated actions are consistent with the goals and regulations of the organization. In a sophisticated phishing scheme scenario, where artificial intelligence cannot detect the email as harmful with high probability, it is crucial for human security experts to carefully examine the identified messages before proceeding with any actions. This joint effort guarantees that important factors, like company objectives and individual business requirements, are considered when making decisions. The ethical considerations surrounding AI in cybersecurity are broad, requiring organizations to address these concerns when integrating AI technologies into their systems. Moreover, the moral considerations of Artificial Intelligence in the field of cybersecurity privacy is an issue when it comes to AI technology, as it relies heavily on collecting and using large amounts of data to work well. This can make people worried about their personally identifiable information (PII) being exposed. Organizations must handle data in a way that follows privacy regulations and ethical guidelines by collecting it and storing it securely. When it comes to bias and fairness in the AI systems mentioned previously, there is a concern that biases from the training data could result in bad results. With that, it is imperative for organizations to proactively ensure that their AI models are fair and inclusive. To do so, they must conduct audits to detect biases and modify them to align with a range of datasets. Organizations need to be open and honest about the ways AI is applied in cybersecurity and have systems in place for taking responsibility when necessary actions are taken by AI that result in negative outcomes.

3.17 Ethical dilemmas and policy issues

As the application of AI in cybersecurity is on the rise, ethical issues and policy challenges are more significant than ever. This chapter also aims to touch on the emerging international laws and recommendations that regulate the use of AI, including those concerning data privacy, fairness, and accountability.

With the continual progress of AI technologies comes the implementation of regulations by governments and regulatory bodies to manage the risks linked with AI development.

The regulatory considerations associated with the use of artificial intelligence for cyber protection are still quickly evolving. It is reasonable, however, to recognize that AI usage will neither change nor remove the requirement to meet applicable regulatory controls and frameworks. Therefore, where a requirement might stipulate something like adherence to Segregation of Duty in the context of Sarbanes Oxley, then

usage of AI will have to adhere to the principle, just like any other aspect of functionality supported in the enterprise.

By way of background, ***Sarbanes Oxley*** (known as SOX to many) is a legal mandate that public companies establish and maintain a robust internal control framework to ensure the accuracy and reliability of financial reporting, often using security frameworks like COSO (Committee of Sponsoring Organizations) as a guide.

GDPR and AI: Ensuring data privacy, the General Data Protection Regulation of the European Union (GDPR) is one of the most significant data protection regulations that establish requirements on the handling of personal data, including the deployment of AI in decision-making processes. All organizations that are implementing AI must ensure that they are compliant with these regulations and thereby give consumers the right to know how their data is being used and to challenge the predictions made by the AI systems on them.

To align with these data-focused regulations, data processing organizations utilizing AI in cybersecurity should guarantee that their AI systems adhere to these rules. This would allow administrators of these systems the option to challenge decisions made by AI systems. For example, the EU Artificial Intelligence Act categorizes AI systems according to their risk levels and imposes stringent regulations for high-risk systems to prioritize transparency and accountability while safeguarding fundamental rights and ensuring fair operation.

The National Institute of Standards and Technology (NIST) has also created recommendations for incorporating intelligence into cybersecurity practices in the NIST AI framework. It looks at transparency, fairness, accountability, and risk reduction in AI systems.

To reiterate, usage of Artificial Intelligence will have to adhere to these principles, just like any other aspect of functionality supported within a regulated company. That is the current state of the practice for AI usage, but the community understands that AI security regulations will continue to evolve toward more established guardrail requirements (NIST, 2022).

3.18 Utilizing AI to improve security resilience in cyber platforms

As companies integrate AI more deeply into their cybersecurity strategies, they can enhance system resilience, ensuring not only strong protection but also the ability to adapt effectively to evolving threats. Building resilience in cybersecurity systems involves establishing an adaptable defensive stance that can swiftly bounce back from attacks while reducing downtime and sustaining operations when confronted with advanced cyber threats.

3.19 Restoring operations after incidents with the help of AI

Ensuring system durability is key, as AI technology has an impact on expediting the recovery process after security breaches occur that endanger systems' integrity. Automated recovery solutions powered by AI can assist in reinstating operations without the need for manual involvement, a process that might be inadequately slow during severe cyber-attacks. As an illustration of this capability, AI can automatically recover data and services by utilizing set backups or duplicated systems to maintain minimal disruptions to business functions.

3.19.1 Case Study: How a Cloud Service Provider Uses Artificial Intelligence for Incident Response

In 2024, a top global cloud provider was subjected to a sophisticated AI-fueled attack that targeted misconfigured rules in its security automation tools. The company had incident response plans that were backed by AI, which isolated the impacted systems and rerouted the network traffic to other servers. Moreover, the AI tools initiated a probe into the cause of the incident that reduced the time for recovery. This resulted in the company having a material improvement in their recovery time while ensuring that there was no adverse effect on their customer service functions (Verizon, 2024).

The use of self-healing systems based on AI can preserve the integrity of cybersecurity defenses, for instance, firewalls and externally facing web-based applications, when they are under attack. Through the use of self-healing mechanisms that can identify and rectify vulnerabilities, businesses can safeguard their infrastructure and services. These systems are particularly useful in the finance, healthcare, and online retail industries that are most vulnerable to disruptions and breaches that could lead to losses and damage to the image of the company. The use of artificial intelligence in the detection and response to cybersecurity incidents is a significant advancement that could shape future trends in the field. The continuous evolution of AI technologies and their integration into cybersecurity protocols will play a critical role in identifying and mitigating emerging threats. The precise and automated nature of AI-based detection mechanisms will lead to more efficient and timely responses, reducing the impact of incidents on businesses. However, there will be challenges, such as the potential of AI tools to make inaccurate decisions, leading to false negative or positive alerts. There is a need for continuous training and fine-tuning of AI models to ensure their effectiveness in detecting new patterns of behavior. Additionally, the ethical use of AI in cybersecurity is crucial, as unfairness or bias in AI algorithms could disproportionately affect certain regions or individuals. In conclusion, the adoption of AI in cybersecurity, specifically in incident response mechanisms, will become a foun-

dational practice for organizations in the future. The benefits of reducing recovery time, maintaining service delivery, and enhancing overall security posture will drive this shift. As AI continues to advance, the boundary between technology and cybersecurity will blur, creating new opportunities for defense but also new challenges for attackers.

3.20 Continuous threat intelligence

In the era of vast artificial intelligence advancements, developing strategies that enable real-time monitoring of cybersecurity systems is crucial for promptly and effectively mitigating risks. For instance, security teams can address risks more effectively by using AI-powered threat intelligence platforms. These systems utilize machine learning to gather and analyze extensive data from various sources, like external threat databases and network traffic, as well as dark web activity, to identify trends and offer practical insights that enhance resilience against cyber threats. Building a flexible security stance, AI systems can adapt, which helps enhance security measures by being more flexible and responsive to evolving threats in the cyber landscape. In the fight against zero-day vulnerabilities, where security gaps are exploited by cybercriminals before fixes are released, AI systems prove themselves most valuable due to their ability to continuously and independently refine their defense tactics. This autonomy to learn and adjust to emerging attack methods empowers organizations to bolster their security posture swiftly and effectively without relying on interventions or awaiting software updates. AI involvement in cybersecurity is exemplified by the implementation of AI-enhanced endpoint security measures. In cases where an AI system identifies behavior patterns on an endpoint (i.e., a surge in data transmission or the execution of unfamiliar applications), it can initiate an immediate lockdown on that endpoint to halt any potential malicious actions. These adaptive security protocols enable organizations to confine cyber-attacks and minimize their repercussions.

For instance, if an employee attempts to access resources beyond their normal access patterns, the AI system promptly flags such unusual activity for further investigation or temporarily restricts access to prevent potential threats. Behavioral analytics enhanced with AI technology assist organizations in gaining insights into user behaviors and enhancing preemptive security measures against insider threats and sophisticated cyberattacks like advanced persistent threats. Through analysis of user interactions and behavior patterns learned over time by AI algorithms, malicious activities such as impersonation of valid users or unauthorized use of compromised credentials can be quickly detected.

Other security systems, such as Security Information and Event Management (SIEM), are now being powered by artificial intelligence technology. This smart system monitors user activities on devices and detects any unusual behavior instantly. For

instance, when a user logs in from a strange location or at an unusual time, the AI is capable of taking automated actions like locking accounts or requesting authentication steps to stop access. AI not only helps in recognizing and stopping dangers but also aids in automating incident response and system restoration procedures efficiently and effectively once a breach is detected.

3.21 Automated fixing

AI technology can be utilized to implement updates or trigger predetermined recovery measures. For instance, once it identifies a software infiltration, the AI system may automatically disconnect the compromised device from the network, reverting the system to its original condition, or install a security update if needed.

AI-driven self-repair mechanisms are gaining popularity in systems, as they can automatically address security issues upon detection without intervention. In a healthcare setting, for example, an artificial intelligence-powered network intrusion detection system can discover a security breach involving the spread of software across the hospital network before it impacts critical life-saving systems. The AI system can detect the unusual activity but can also take swift action by isolating the malware to prevent any impact on essential patient care systems. It can then activate an automated recovery procedure to fix any affected systems quickly, minimizing downtime and ensuring uninterrupted patient care services.

Organizations can enhance their resilience by leveraging AI-powered analytics to anticipate threats using past data and sophisticated machine learning technologies. AI can gather threat intelligence from channels like established attack repositories and information from online platforms and communities that exchange threat data. This data is valuable for creating models that anticipate hacking techniques or weaknesses that hackers could capitalize on.

Early alert systems use AI-driven algorithms to detect security threats by identifying early signs of suspicious activity before they escalate. These platforms monitor hacker behavior, such as probing for system vulnerabilities or attempting network infiltration, allowing for proactive defense and swift response measures to prevent security breaches.

This technology can assist governmental organizations that protect critical infrastructure. By utilizing an AI system to examine up-to-the-minute threat data, the security system can successfully anticipate a cyber assault on a nuclear power facility by detecting recurring patterns in the conduct of hackers who had previously used these techniques in other recorded attacks. By issuing alerts, an agency can proactively implement security measures to prevent and lessen the threat.

“Adjustable Security Stance”: The learning and adaptive capabilities of AI contribute to boosting resilience by enabling systems to modify security protocols in response

to the changing threat landscape. This adaptability ensures that security measures remain efficient amidst evolving attacker strategies, further underscoring the value of AI.

"Adaptive Access Controls": Utilize AI to modify access levels in response to real-time threat intelligence findings. For example, when an AI system identifies an increase in threat activity, it can enhance access restrictions by implementing authentication methods (like requesting two-factor authentication or limiting privileges). This is done to protect data and assets effectively.

"Dynamic Risk Assessment": Involves AI models that analyze network conditions and behavior patterns to evaluate risks with real-time precision. Through user activity monitoring, security measures are adjusted to address threats when high-risk users or devices display suspicious activity or become targets of evolving attacks. In a company setting, security measures are regularly updated by AI according to the threat levels specific to a region. Some regions are considered to be in high-risk zones where cyberattacks occur frequently. The AI system enforces security protocols accordingly, like demanding authentication at shorter intervals and implementing network segmentation to limit access to sensitive systems. In today's fast-paced organizations, data is increasing at a rate that poses a challenge for conventional monitoring systems to respond promptly to potential risks. The advantage of AI systems lies in their ability to continuously monitor networks, user activities, and applications at a scale while processing vast amounts of data in real time.

AI-powered security tools offer real-time threat detection by analyzing extensive data streams to detect attacks efficiently, in contrast to human-operated security operations centers (SOCs). This continuous monitoring capability enables the identification of threats that traditional security systems may overlook; these include zero-day attacks and polymorphic malware, as well as insider threats.

"Automated Data Correlation": Enhances AI's capability to connect data from security sensors and logs, effectively spotting attack sequences that could otherwise slip under the radar. By amalgamating information from firewalls, intrusion detection systems, endpoint security tools, and other outlets, AI can construct a perspective of the network's security status. This promotes effective decision-making and quicker identification of advanced threats.

Let's play out an example of a prominent technology corporation that leverages artificial intelligence to oversee its infrastructure. By examining countless data points every second to identify irregularities, such as sudden spikes in traffic or potential credential stuffing attacks, the AI system is programmed to alert or respond defensively when abnormal behavior is detected. The aim or end objective is continually safeguarding against cyber threats (24x7) without an overwhelming need for large teams of skilled security personnel. Whether a large company or government entity, organizations do not have infinite budgets or access to the number of highly skilled personnel to take on the job at hand. With that, AI is essential in leveling the playing field for defenders.

3.22 Pros and cons of using AI in cybersecurity

Although AI is very useful in the field of cybersecurity, several problems and limitations come with the use of AI that organizations should know and work on. Sometimes, the complexity of the AI systems can lead to the development of some vulnerabilities or problems in the usage of the systems.

3.23 High cost of implementation and integration

The integration of AI into cybersecurity functions usually demands a high level of investment. This includes the cost of buying AI tools and platforms but also the costs of training and updating the AI models. For most small businesses or companies with limited cybersecurity budgets, these costs are out of reach. Nevertheless, the expense of implementing AI will eventually come down and is already being offered by third-party providers in the form of 'cloud-based' cybersecurity offerings. These providers and the respective solutions they are offering will also do much of the integration legwork as they partner with well-known technology software and platforms. In the end, this will result in positive gains in the overall costs to defend against cyber-attacks in the future because AI systems help organizations avoid breaches and improve on the time it takes to respond to an incident, thus avoiding incurring huge financial and reputation losses.

In any case, an Artificial Intelligence-based cybersecurity system will need to integrate with two broad types of technology. First, the AI system must have visibility to the systems it is looking to protect, including a collection of telemetry for visibility and the ability to mitigate risk in an in-line fashion (i.e., the ability to perform actions on target systems). Secondly, the AI cybersecurity system must integrate with other security technologies across the enterprise. Security information event management (SIEM), for example, will need connectors from the AI cybersecurity solution to ensure that any relevant alerts and alarms from the system are normalized with other security information.

Moreover, an AI cybersecurity system must integrate with more conventional, yet advanced, machine learning operations, which are already built into the Security Operations model and are focused on drawing inferences from the consumption of large amounts of data. With that, machine learning models will need to be streamlined and leveraged to achieve the greatest threat defense success. Integration is key and foundational to successfully leverage AI and make this work as part of an overall cybersecurity architecture.

3.24 AI system security and vulnerability management

AI systems are not shielded from various cyber-attacks as conventional systems. Model inversion attacks, for instance, can reveal the training data of the AI models, thus giving the attacker access to sensitive information. Data poisoning attacks, which have been discussed earlier, are used to manipulate the training data in such a way that the AI systems will make improper decisions. Furthermore, the adversaries may try to reverse engineer the AI models or find vulnerabilities in the algorithms used in AI-based security solutions.

Some AI models are complex and somewhat random, making it hard to understand their behavior, which creates new cyber risks. To prevent these threats, organizations should adopt AI security frameworks that entail testing and validation of the AI models to ensure that they are secure and robust. These security measures can assist in identifying the vulnerabilities and fixing them before they are used.

3.25 Industry specific threats and AI countermeasures

The usage of AI in cybersecurity is entirely inconsistent across various industries, each of which has its set of issues and dangers.

For example, let's look at the retail sector and AI applications used in cybersecurity for E-commerce. The retail industry is also leveraging AI to offer personalized experiences, automate inventory management, and fight fraud. Nevertheless, the application of AI and digital platforms in retail trade opens new opportunities for cybersecurity threats.

An example of an AI application in e-commerce fraud control would involve an AI-based fraud detection system used to analyze the activities of the users and provide real-time results. The system would be able to detect new fraudulent transactions, but at the same time, the retailer would have to be wary of the number of false positives that could lead to a delay of payment processing to legitimate customers and thereby reduce the level of customer satisfaction and possibly the company's revenue.

This creates a challenge of using AI for detecting fraud in retail while upholding the need to support a positive customer experience. To accomplish this, an e-commerce platform would need to ensure that its AI model has levers to include human intervention in the fraud detection process. A team of fraud experts would need to intervene to examine the transactions flagged as suspicious. This co-existence of AI and humans would lead to a lower incidence of false positives while still flagging genuine threats. Companies will need to employ the concept of continual learning in their AI systems to

enable the models to learn from previous experiences and enhance their fraud detection capabilities over time.

3.26 Navigating the future of AI in cybersecurity

AI integration in cybersecurity offers opportunities and risks. On the one hand, AI technologies can bring significant improvements in the protection of organizations' digital domains; on the other hand, they may become new vulnerabilities that can be exploited by the enemy. In recent times, AI has been developing fast, therefore, organizations must have a good risk management plan that addresses these new risks. To this end, a risk-based approach, strong governance, and ethical principles are vital to tackling the challenges of the AI-based cybersecurity environment. Combining AI with human expertise ensures cybersecurity remains adaptive to emerging threats. Through encouraging openness, responsibility, and conformity to the appropriate rules, companies can fully tap into the possibilities of AI while avoiding the risks that come with it. Thus, in the AI era, the dual objective of improving performance and avoiding negative impact is crucial. To this end, organizations can sustain themselves in this new normal of AI security practices and ensure that the role of AI in cybersecurity is beneficial and secure.

3.27 Conclusion

The future outlook for Artificial Intelligence in cybersecurity is promising, as it is hard to imagine any company, government agency, or other organization that will not be using AI more extensively in the coming years. The cybersecurity market continues to evolve and expand as organizations across industries prioritize their security needs and how AI will assist in their overall defense architecture.

To this end, artificial intelligence will likely become an essential component of nation-state defense. As such, we should expect that these entities will drive toward increased use of AI for military and other government applications. This is sensible since nation states are always trying to be on the cutting edge of using the best technologies for offense and defense.

From a solution provider perspective, the chances are also high that these providers will begin to offer managed AI cybersecurity services, especially for smaller companies that might not have budget to properly protect against AI threats, both internally (i.e. the inherent risks of using AI for business purposes) and externally (i.e. the risks of attackers using AI to breach the organizations defenses). This also makes sense since AI involves automation, which scales nicely for managed services.

The integration of AI cybersecurity into platforms such as Secure Access Service Edge (SASE) and associated cloud-managed security services is high, as the new AI integrations into cyber products and services continue to demonstrate the viability of a market for this type of AI protection.

As we continue to move further into the future of AI, cybersecurity professionals must continue to work vigilantly, flexibly, and preemptively. The integration of AI in the cybersecurity domain opens great possibilities for both the identification and prevention of cyber threats. However, there are new risks that come with it, and they must be properly addressed. Organizations must take a systemic approach to AI cybersecurity, optimizing the benefits of AI technologies while ensuring that human capital and ethical principles are not neglected. With awareness of the risks, investment in AI security, and compliance with evolving regulatory standards, companies can effectively steer their way through the challenges of protecting their digital assets. In the future, the development of AI cybersecurity will be based on cooperation between AI solutions and human experts and between different industries sharing experiences and findings. Thus, building a resilient, ethical, and secure AI ecosystem will enable us to reap the benefits of AI with minimal risks and build better digital communities for everyone.

Using AI in improving the cyber resilience of organizations' systems is not only about preventing cyber-attacks but also about getting ready for them, responding to them, and recovering from them in an automated, intelligent, and optimal manner. AI can be incorporated into various aspects of an organization's cybersecurity function, including threat detection, incident response, predictive analytics, and continuous monitoring to develop systems that are not only secure but also resistant to the many complex threats that exist in the current world. The effectiveness of AI is evident in its capability to learn and adapt to the dynamic threat environment and to perform menial tasks that enhance the efficiency of the cybersecurity operations center. With the help of AI, cybersecurity systems are in a better position to absorb and recover from a breach, thus ensuring the continuity of business operations, the protection of important information, and the trust of customers and other stakeholders.

As AI technology develops, the companies that decide to build their cyber resilience on AI will be in a better position to deal with future cybersecurity risks and innovate their approaches to address new threats.

References

Gartner, 2023. Predicts 2024: AI & Cybersecurity — Turning Disruption Into an Opportunity (WWW Document) URL https://www.gartner.com/document-reader/document/4995731?ref=solrAll&refval=470573646 (accessed 3/5/2025).

CrowdStrike, 2025. CrowdStrike 2025 Global Threat Report: Adversaries have adapted. Have you?) (WWW Document) URL https://www.crowdstrike.com/en-us/global-threat-report/ (accessed 3/6/2025)

FBI, 2024. FBI Warns of Increasing Threat of Cyber Criminals Utilizing Artificial Intelligence, FBI, (WWW Document) URL https://www.fbi.gov/contact-us/field-offices/sanfrancisco/news/fbi-warns-of-increasing-threat-of-cyber-criminals-utilizing-artificial-intelligence (accessed 3/12/2025)

NIST, 2022. AI Risk Management Framework (WWW) URL https://www.nist.gov/itl/ai-risk-management-framework (accessed 3/1/2025).

Verizon, 2024. Data Breach Investigations Report (WWW Document) URL https://www.verizon.com/business/resources/reports/2024-dbir-data-breach-investigations-report.pdf (accessed 3/19/2025).

TechTarget, 2024. How AI is making phishing attacks more dangerous, TechTarget (WWW Document) URL https://www.techtarget.com/searchsecurity/tip/Generative-AI-is-making-phishing-attacks-more-dangerous (accessed 3/4/2025).

CrowdStrike, 2023 Machine Learning (ML) in Cybersecurity: Use Cases, CrowdStrike (WWW Document) URL https://www.crowdstrike.com/en-us/cybersecurity-101/artificial-intelligence/machine-learning/ (accessed 3/7/2025).

Muhammad Z. Hussain, Muhammad Z. Hasan and
Muhammad R. Siddique

4 Mitigating risks in AI-driven network security: Strategies for ensuring resilience and integrity

4.1 Introduction

Artificial Intelligence (AI) is transforming network security by enabling organizations to detect, predict, and respond to cyber threats in real time. Traditional rule-based systems are limited in dynamic threat environments, whereas AI-driven security solutions offer automation, scalability, and advanced pattern recognition. These systems can analyse gigabytes of network traffic data to distinguish between normal and suspicious behavior, predicting anomalies before they escalate into threats. Unlike conventional security systems that rely on known attack signatures, AI models learn from historical patterns and adapt to new attack strategies, making them especially effective against advanced persistent threats (APTs), zero-day exploits, and sophisticated malware (Sarker, Furhad, & Nowrozy, 2021).

One of AI's most important contributions is automation. AI-powered tools can monitor network activity and autonomously respond to irregularities without human intervention. This includes isolating infected devices, blocking malicious IP addresses, and preventing the spread of threats across networks. Automation significantly reduces response times, minimizes damage, and ensures business continuity (Bellamkonda, 2024).

AI also enhances behavioral analytics and anomaly detection. As attack methods rapidly evolve, static threat intelligence is often inadequate. AI-based behavioral models establish baselines of normal activity and flag deviations—such as unusual login times, data transfers, or unauthorized access—as potential breaches. This proactive approach aids in detecting insider threats, privilege escalation, and account takeovers that traditional systems might miss. In addition, AI can aggregate and interpret threat intelligence from a variety of sources, including the dark web and open-source intelligence (OSINT), helping cybersecurity professionals to adjust defensive strategies accordingly.

However, there are significant risks associated with using AI-driven security, one of which includes adversarial attacks. This is when cybercriminals insert false data into AI training sets in order to alter model results. An AI model that is trained on corrupted data can fail to identify actual attacks. AI models often require access to large datasets, including confidential credentials and proprietary organizational data. These are easy targets if appropriate data governance, encryption, and access restrictions are not in place.

https://doi.org/10.1515/9783111595306-004

Cybersecurity analysts need to continue to validate AI-generated alerts and adjust systems to ensure appropriate responses, being cognizant of false negatives and false positives. A false negative is the failure to identify a genuine threat, whereas a false positive incorrectly identifies benign situations as risks. This may cause security teams to receive too many pointless notifications. AI plays a critical role in modern cybersecurity because of its ability to process large-scale data, detect complex attack patterns, and respond rapidly. Nevertheless, organizations must manage the risks of AI adoption by refining models continuously and maintaining human oversight. The most resilient approach combines AI-driven automation with human expertise in a layered, adaptive defense strategy. AI is essential due to its capacity to handle vast amounts of data, identify intricate attack patterns, and react quickly. Organizations must implement AI carefully, using a multi-layered, adaptive security system that blends human experience with AI-driven automation. AI will become central to protecting digital assets, infrastructure, and sensitive information as cyberspace becomes more complex (Salem et al., 2024).

4.2 AI's role in enhancing network security

AI is revolutionizing network security by equipping security operators with tools that enable broader and faster protection. As cyber threats grow in complexity and frequency, traditional security systems are proving to be insufficient. AI-powered security solutions—featuring deep learning, machine learning, and data analytics—enable automated, real-time threat detection and predictive defense, helping organizations stay ahead of emerging risks (Salem et al., 2024).

AI offers significant value by processing vast datasets from network logs, endpoints, and threat intelligence feeds. Unlike human-operated detection, which struggles with time and scale, AI systems quickly analyze anomalies using machine learning models which identify patterns of normal and abnormal activity, enabling early detection and response to zero-day attacks and APTs. The rising threat levels make rule-based systems and signature detection tools inadequate; however AI's behavioural and anomaly detection techniques are far more adaptive and are able to respond adequately (Umoga et al., 2024).

Rapid detection and response are essential during breaches, and a critical advantage of AI in security is its speed. This is particularly crucial in ransomware attacks in which time is of the essence. With AI, security teams gain the ability to autonomously isolate endpoints, block malicious IPs, and implement automated countermeasures, reducing encryption risks, and preserving business continuity. With the use of AI, a security team can implement predictive analytics and behavioural analytics. These models are used to anticipate attacks by analyzing past incidents, enabling proactive actions like patching vulnerabilities and updating firewalls before attackers strike.

Likewise, models can be used to detect insider threats and compromised credentials in real time.by screening for unusual behavior, such as unauthorized logins or abnormal data transfers.

Deception techniques are also enhanced by AI, including the use of honeypots and decoy networks. that simulate legitimate environments. These methods lure attackers to study their tactics and help divert attackers away from sensitive assets to less critical systems (Dasaratha Rao et al., 2024). Nevertheless, AI security systems face challenges as attackers also use AI to assist in their efforts to generate deceptive emails, perpetrate phishing schemes, clone voices, and create deepfake videos. Poorquality or biased training data often leads to flawed models that misclassify threats or allow harmful activity to go undetected, and, adversarial attacks, in which bad actors confuse the system by manipulating AI inputs to remain a serious risk. While AI automates threat detection, prediction, and response, human oversight is still crucial and security analysts must validate AI outputs, guide model updates, and ensure ethical use. AI is indispensable for protecting digital assets, infrastructure, and sensitive information relying on its real-time capabilities, predictive power, and adaptive defenses to make it central to modern cybersecurity strategy—provided its risks are managed with care and its deployment is guided by human expertise.

4.3 Risks and vulnerabilities introduced by AI

Networks have been primarily secured, and AI has made remarkable progress; however, risks remain, and there are vulnerabilities that organisations must take care of. As AI-driven security systems increase, criminals are also getting more sophisticated about AI model exploitation. While AI excels in processing vast datasets and recognizing patterns, its architecture is susceptible to manipulation. The five most critical risks associated with AI on network security are adversary attacks, data poisoning attacks, model bias, lack of transparency, and overreliance on automation.

Adversarial attacks are among the most ominous security threats. These attacks constitute cyber criminals posting specially crafted inputs to the AI model causing misclassification. Malware samples can be slightly altered, and being imperceptible to human analysts, evade detection by security systems. For that reason, just like a hacker eluding the human perception, these hacks have been able to comprise minute changes so that the changes are seemingly invisible to the human eye. These manipulations exploit the mathematical foundations of machine learning algorithms to bypass defenses and gain unauthorized access.

Data poisoning is another major attack when bad actors influence the dataset being fed into the training of the AI models. Data is at the very core of an artificial intelligence system. Therefore, if we put biased data in a training set, we can say that most of the accuracy and dependability of an AI model will become useless. For

instance, a dataset used for training an intrusion detection model could be corrupted, and the pattern of legitimate traffic and non-legitimate traffic is confused. Poisoned datasets compromise AI's core functionality and can be difficult to detect or reversed once embedded in the system.

Model bias is another primary concern regarding how we rely on AI network security, and emerges when AI systems are trained on unrepresentative or imbalanced datasets. If a model is exposed to a narrow range of data, it may perform poorly when encountering unfamiliar threats, particularly those outside the training scope in terms of geography, industry, or attack type. This issue is especially challenging in fraud detection, access control, and automated surveillance systems, where bias can lead to unreliable outcomes.

A related matter is the lack of transparency in many AI-driven security systems, which often appear as "black boxes," that offer little insight into how decisions are made. Analysts receive alerts without understanding the underlying logic and this opacity impedes auditing, error diagnosis, and model refinement, creating opportunities for attackers exploit. Utilizing explainable AI (XAI) techniques help address this problem by making AI decisions more interpretable for human analysts.

While AI systems can reduce analyst workloads and response times, they are not foolproof. AI automated security solutions may not detect complex attacks that are not in the training pattern, and human oversight, an organisation may miss significant threats, or be ill-equipped to respond to occurrences to which AI systems are prone. Additionally, there are privacy issues associated with AI-led security systems, which process vast quantities of sensitive data, such as user behavior logs and access credentials. These systems are an enticing target for cybercriminals to extract confidential data and the use of AI to explore and watch over the most sensitive data can infringe on user privacy, particularly if deployed without transparency or consent. To comply with regulations like the General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA), organizations must implement strong encryption, access controls, and data anonymization strategies.

Adversarial attacks, data poisoning, model bias, opacity, and over-automation all pose significant threats, which can be mitigated by organizations that prudently invest in adversarial defense, ensure diverse and representative training data, incorporate explainable AI techniques, and maintain strong human oversight. By taking a balanced approach, it is possible to leverage AI's strengths while minimizing its vulnerabilities and preserving both security and trust in digital systems.

4.4 Risk identification

It is important to identify risks in an AI driven security system to ensure trust. AI security models are vulnerable to attacks by adversaries if not thoroughly assessed and

controlled. Issues can sometimes be traced back to modelling problems, whether stemming from the training data itself or through ambiguous manipulation techniques. Organisations need a streamlined systematic process for identifying and evaluating AI driven security solutions to keep a strong security posture.

A cyber-criminal can deploy an AI based intrusion detection system (IDS) and craft an adversarial input which the AI based IDS marks as benign. Since AI models depend on large datasets on which the models make their decisions, AI models are also good targets for cyber criminals. Misconfigured AI systems could end up leaking confidential network logs, access credentials, and behavioural analytics data. Considering these risks, we need to mitigate them and use some privacy preserving AI techniques like federated learning and differential privacy, to allow the AI models to study data in a way that has sensitive information hidden. A major problem is if an AI security model is developed using a biased or non-representative dataset. In such situations AI-based security will result in false positives or false negatives and damage to the system's reliability

Another risk factor is that of the efficacy of AI driven security models via concept drift. Older trained AIs models also pose risks, because cyber threats are evolving and cybercriminals constantly do things in different ways. Consequently, organizations need to create continuous learning frameworks that encourage continuous retraining of its AI security models with periodically retraining them using current threat intelligence so they can always know what to do when receiving the next threat alert. More importantly, such automated model monitoring systems should help automatically measure performance for detecting performance degradations and update performance when necessary.

Are these risks effectively identified by organizations? That is why the organization should apply a structured risk assessment methodology containing risk monitoring, penetration testing to be implemented regularly, and AI security audits. Other explainability and interpretability techniques of giving insights on the AI decision making process include Shapley Additive Explanations (SHAP) and Local Interpretable Model Agnostic Explanations (LIME). Additionally, these provide security teams with insights into why an AI model labels a particular thing as malicious or not.

The evolution of AI driven security model has certainly changed the face of the cybersecurity completely. However, these also bring along several inherent risks for cyber criminals to exploit. It is well known these risks include adversarial attacks, model bias, and data privacy concerns, putting into question the integrity, accuracy, and security of AI powered systems. An important first step to develop reliable mitigation strategies for AI driven security solutions is to understand these risks.

4.4.1 Adversarial attacks

Adversarial attacks are the biggest threat to the AI based network security. Such attacks are performed in which an attacker deliberately changes input data resulting in the AI model making a wrong decision or misclassifying an object. The cybercriminal then initiates an attack which the AI model labels as non–threatening when it really is threatening, or the other way around, as a suspicious input when it is not. For example, a method to bypass a given AI antivirus program is by introducing or removing a huge amount of noise in an executable file or rearrange its code to meet the same function. In biometric authentication systems using facial recognition, for example, counterparts to adversarial perturbations can allow unwelcome persons to circumvent security barriers.

Adversarial attacks make use of mathematical dependencies of AI models so inherently that the weaknesses found in how feature extraction and the classification mechanism of the AI models are achieved act as the cornerstone for adversarial attacks. Different attack forms include evasion attacks wherein attackers change malware signatures to evade detection, poison attack wherein attackers insert corrupted data in the data fed to the AI training set. Adversarial vulnerabilities to AI can greatly reduce effectiveness of security measures involving AI.

4.4.2 Model bias

Model bias is one of the most serious risks in an AI based cybersecurity system. The data we have from history does not come in neat, clean, perfect sets. It can be incomplete, imbalanced, or skewed, and when training an AI model using such data, we end up with unfair, inaccurate or even discriminatory decisions. AI security models that are biased can trigger false alarms, and miss some of the attack vectors. For instance, if an AI fraud detection system is trained on one specific geographic area or one specific customer demographic, then the effectiveness will fail when it deals with fraud activity specific to a different geographic area or different customer demographic. In a similar way, if a model trained to discover intrusions over the network with enterprise network attacks, but not those caused by the Internet of Things (IoT) attacks, will not be able to identify cyber threats of the IoT devices. When the AI system is learning from a biased training set, model bias can produce security gaps because the AI system cannot find new possible threats that fall outside its training set. Also, AI based security decisions are liable to ethical or legal issues if not sourced from a balanced population and hence their treatment of some communities can be perceived as unfair.

4.4.3 Data privacy concerns

AI models need access to security related data, including detailed network logs, user behaviour patterns, and access credentials. If proper security is not implemented, then data that is linked with it can be a potential target to a cybercriminal who will try to gain access to this AI driven system. Without effective data governance and protection mechanisms, large sets of data are collected and stored unsecured posing a high risk for the organizations. If we train on the data with unprotected or incorrect anonymization, we can expose confidential information. In addition, AI models have occasionally 'memorized' details from training data and inflicted data 'leakage' – the circumstance of the model divulging some sensitive data.

4.5 Techniques for identifying risks specific to AI models used in network security

Proactively, one has to identify and mitigate risks in AI driven network security models to avoid a deficiency in reliability and security. As with any technology, AI security models are subject to attacks, biases and system wide vulnerabilities. Thus, organizations need to implement effective risk assessment methodologies towards such as: threat modelling, adversarial testing, bias audits, explainability techniques and continuous monitoring.

4.5.1 Threat modelling for AI security risks

Threat modelling is a structured approach to identifying potential security vulnerabilities in AI systems before they can be exploited. In AI-driven network security, threat modelling assesses the attack surface of AI models, considering all potential points of compromise, such as adversarial manipulation, data leakage, and model inversion attacks. Threat modelling provides a proactive framework for identifying AI security risks before attackers can exploit them, ensuring AI systems remain resilient against evolving cyber threats.

4.5.2 Adversarial testing and red teaming

One of the most effective ways to defend against adversarial attacks is through adversarial training, where AI models are deliberately exposed to manipulated inputs during their training phase. This testing allows organizations to know how AI systems behave under hostile circumstances and locate areas where attackers can take advan-

tage of. Red teaming uses the help of ethical hackers, or AI security specialists mimicking adversarial attacks, against AI driven security tools. Red teaming gives a deeper view to the resiliency of an AI model in an actual world attack tactic.

4.5.3 Bias audits and fairness testing

While AI security models might prove biased, they could incorrectly alarm some types of attacks, or specific groups of users. The AI security models are systematically audited to find out when the system is making unjustified decision patterns. Fairness testing is especially critical in AI-powered fraud detection, biometric authentication, and access control, where biased models can result in discriminatory security measures.

4.5.4 Explainability methods

Many of today's AI driven security systems are a 'black box,' deciding actions without explanations of how and why they made such decisions. Security professionals use explainability methods to interpret the AI-driven security decisions and point out the problems of the AI models. In the case of AI driven threat detection, fraud prevention and anomaly detection, explainability is very important function as security teams need to be able to validate and act upon the decisions of an AI model, and to be able to understand why a certain event was flagged as anomalous or fraudulent.

4.5.5 Continuous monitoring and real-time anomaly detection

The effectiveness of AI models in combating new cyber threats is effective only when being constantly monitored. Attacks change at a very fast pace and effective security tools require AI responding in real time to new attacks. By enabling the execution of real time anomaly discovery, organizations attempt to ensure their AI security models are accurate and current to do their job of mitigating new cyber threats.

Adversarial testing helps to understand the external behaviour of the AI model in hostile environments, whereas threat modelling can be used for understanding system vulnerabilities. Bias audits and fairness testing are used to maintain the accuracy and the fairness of the AI security decisions. AI security defense can enhance anomaly detection in real time and provide permanent supervision to make sure the models are updated with new and changing cyber threats. Used as organizational cybersecurity techniques, these techniques can diminish the AI security risks by rendering the use of the AI to be secure.

4.6 Real-world cases

The penetration of AI in network security is such that it is constantly bringing about a new dimension in identifying and mitigating threats. However, AI security systems deployed in the real world can be attacked, have bias, and can be exploited. AI models are being recruited by cybercriminals, are being tricked into training themselves and then managing to bypass security controls and infiltrate digital infrastructures, defeating security controls. The following case studies present the vulnerabilities of security systems that use AI and the central aspect of having proper risk management strategies.

4.6.1 Microsoft's tay chatbot: adversarial manipulation of AI

A well-known case of how AI can be misled by adversarial manipulations is Microsoft's Tay chatbot. In 2016, Tay was started as an AI interactive chatbot that would teach itself from human interactions on Twitter. As Microsoft stated, Tay was supposed to evolve in conversation with users and learn language processing skills. However, Tay was fed harassing, offensive, and blatantly political statements by malicious users. Tay was using unsupervised learning and started to absorb the toxic input and tweet racist, misogynistic and inflammatory content. The serious flaw in AI driven learning systems is that they are vulnerable to adversarial manipulation and data poisoning attacks. The attackers modified the self-learning algorithm of the chatbot by feeding it with malicious data that poisoned its ability to produce appropriate responses. Also, the inability to apply proper precautionary measures with the ethical AI forced it to be turned offline within 24 hours as it became necessary to have real time moderation and adversarial defense to stop the AI from behaving badly.

4.6.2 Malware detection systems bypassed by adversarial techniques

The idea here is to detect and neutralize threats by looking at the software functionality and file signatures. Despite that, such AI based systems have been attacked through minor modification of malware code, referred to as adversarial evasion. A decade ago, research showed that with simple changes to the malware sample, they could completely befuddle AI powered AV engines. Without impacting functionality, attackers could make ever so slight changes to the structure of the malicious code that the AIs could not discern. As these AI security models rely on the feature extraction, they can only work using the features for this case since other features cannot be extracted. On the other hand, if the AI model is trained based on the historical attack data, they may fail to recognize the new refused threats in its new form that is differ-

ent than what it has been taught to learn. However, security experts suggest that adversarial training is the solution to this; introduce AI models with maliciously adversary altered malware samples to make them more resilient to evasion attacks. Along with that, they can be used to make hybrid security models which will work on rule-based detection of malware as well as AI detection to fortify the defense against growing cyber threats.

4.6.3 Deepfake phishing attacks: AI-generated social engineering

Cybercriminals use AI machine learning based deepfake tools, which have become powerful weapons in the world of phishing attacks and fraud. In 2019, cybercriminals famously targeted an employee at a UK energy firm with AI deepfake audio mimicking the company CEO to obtain $243,000 from an employee in a fake bank account. The phishing attempt was very convincing, as the attackers took AI-driven voice synthesis technology and rendered the executive's voice patterns, tone, and accent using it. In this case, social engineering attacks using AI are becoming more and more dangerous. Deepfake phishing attacks differ from regular phishing attacks that trick humans through fraudulent emails or websites as they involve AI generated content to fool human senses. The problem is that these are particularly dangerous, since traditional cybersecurity measures like email filters and network monitoring are not included in detecting them. Deepfake phishing is one that requires organizations to utilize multi-factor authentication (MFA), anomaly detection, and voiceprint verification to combat the phishing since these are risky communications and phishers need to ensure authenticity.

4.6.4 Lessons learned and risk mitigation strategies

The above case studies reveal key insights into the vulnerabilities of AI-driven security systems and highlight critical risk mitigation strategies, including: AI models must be designed with robust adversarial defences; continuous monitoring and human oversight are essential; advanced authentication methods are necessary to prevent AI-powered social engineering; and AI models must be continuously updated to adapt to emerging threats.

Events that took place have proved that while indeed AI makes the cybersecurity better than before, it also presents another side of the coin. We observe adverse manipulation and poisoning of data in the Microsoft Tay chatbot fiasco. Furthermore, this attack showed that in one of the malware domain detection bypass cases, adversarial evasion risk is a real attack risk when attackers use malicious code to bypass the security AI tool. The technique demonstrated the increasing peril to AI social engineering using synthetic media to fool human perception.

Because such technologies need protecting from threats like these, organizations using AI powered security systems must perform adversarial testing, continuous monitoring, use of explainable techniques, and rigorous authentication modes in order to ensure the reliability of the systems. In order for AI security to be trusted and fair, it needs to be able to adapt to new cyber threats as they appear as well as implement some human oversight and ethical AI frameworks. Proper plans on risk management when addressing AI for cybersecurity in organization will therefore strike the right balance between the advantages and the corresponding risks of AI in cybersecurity.

4.7 Mitigation strategies for AI risks

The adoption of AI-powered security systems offers multiple possibilities by which different risks will arise; these are adversarial scenarios, data privacy issues, bias, and operational issues. In a way, it can be stated that the question of cybersecurity implies the strengthening of the necessary principles and approaches to address risks at multiple levels across an organization. With the implemented strategies on AI model defense, privacy-preserving AI techniques, continuous monitoring, and human supervision, the security teams can further improve the AI-driven security solutions with mitigating risks.

4.7.1 AI model robustness and adversarial defense techniques

As AI becomes ever more vital in the cybersecurity field, adversaries are discovering increasingly sophisticated attack techniques directed at the holes in AI models. The defensive system of adversarial training exposes AI models to altered input data during the training phase to enhance their capabilities in identifying and resisting adversarial attacks. Through adversarial examples that are specifically created to fool AI the model becomes able to recognize and prevent the attacks which attackers make using subtle modifications.

In adversarial training AI security researchers create inputs that are slightly modified but designed to mislead the AI model (e.g., altering a malware signature in a way that makes it look benign). The model is retrained with both normal and adversarial inputs, ensuring it learns to distinguish between legitimate and deceptive patterns. As adversarial tactics evolve, AI models must continuously be updated with new adversarial examples to improve their resilience against emerging threats.

Beyond adversarial training, AI security systems can implement various defensive techniques to strengthen model robustness against manipulation. Input validation and sanitization is done to ensure that AI models process only clean and verified inputs. AI models should reject input anomalies or highly manipulated data points

that deviate significantly from expected values. Noise injection and randomization are used to add random noise, as small variations in the input of training data, to prevent the AI models over reliance on these insignificant variations in input. Model assembling can be achieved by combining different AI models instead of working just with a single model. It means if an adversary exploits one model, other models of the ensemble would still be able to classify the input correctly.

Defensive distillation will make an AI model learn to ignore very minor deflections in its training inputs to make the model less susceptible to adversarial manipulation. Gradient masking camouflages the model's internal computations and thus makes it hard for attackers to trick AI predictions with gradient based attacks. These techniques, and others, are used by organizations so that AI based cybersecurity can remain dependable.

4.7.2 Data governance and privacy-preservation

More and more data is being required by AI driven security systems as they are becoming more sophisticated, and yet, this comes with its own set of privacy and security risks such as unauthorized access to data, data breaches and regulatory noncompliance. Therefore, organizations need to have a strong data governance framework and privacy preserving AI techniques to protect the sensitive data and to secure the security models of AI.

Implementing strict access controls ensures that only authorized personnel can access sensitive data. Role-Based Access Control (RBAC) restricts data access based on job roles and responsibilities, reducing the risk of insider threats. Multi-Factor Authentication (MFA) and Privileged Access Management (PAM) add additional layers of security, ensuring only trusted individuals can interact with AI datasets. Secure data storage solutions, including hardened databases and cloud security policies, help mitigate risks associated with AI data management.

Data anonymization and masking anonymization techniques, such as data hashing, tokenization, and pseudonymization, obscure personally identifiable information (PII) while retaining AI data usability. These data governance practices establish a secure foundation for AI-driven security systems, ensuring privacy, compliance, and data integrity.

Organizations can foster data security and privacy even as AI techniques facilitate the training of machine learning. Federated learning, differential privacy, secure multiparty computation are the useful techniques to this end. Federated learning is a training technique that allows the model to be trained using multiple datasets without sharing these datasets to a central point. While traditional solutions put data in one shared database and store them, the federated learning technique enables the training of AI models at the edge of the device like smartphones, IoT devices, or even in the secure clouds. Differential privacy is a technique that adds mathematical noise to

datasets to prevent AI models from revealing individual data points while still maintaining useful insights. Secure multi-party computation (SMPC) allows multiple parties to compute AI models on shared data without revealing the underlying data.

The continuously growing power of AI driven security solutions will inevitably result in an increase in the risks related to the privacy of sensitive enterprise data, unauthorized access, as well as compliance issues. In response, data governance frameworks and techniques to ensure privacy in AI must be in place to minimize such risks. Integration of these privacy enhancing strategies allows organizations to use the power of AI based security systems without compromising on efficiency and security. In the context where cyber threats and privacy concerns are on the rise, a proper AI governance framework provides a possible solution for the future of AI powered cybersecurity.

4.7.3 Continuous monitoring and anomaly detection

Security analysts need to be vigilant to detect possible network risk because threats are ever-changing. According to the security model type, conventional models come from a rule-based setup and are not adequate in finding new and evolving threats. However, AI security systems using machine learning (ML) and behavioural analysis detection can more readily react to probable security issues in real time. Security threats that could cause losses can be defended by automated anomaly detection and adaptive AI, in addition to regular supervision. AI-driven Security Information and Event Management (SIEM) platforms collect and analyse security logs from various sources to detect threats in real time. Modern SIEM systems incorporate AI and machine learning to automate threat correlation, anomaly detection, and predictive analytics

The process of continuous monitoring entails checking network traffic, system logs and other activities to identify irregularities that may signify threats. In contrast to conventional recovery and check-up approaches of security assessment, it is more effective when threats are spotted when they are arising and hence can be addressed promptly. Continuous monitoring analyses data streams in real time, identifying malicious activity before attackers can gain control. AI-powered monitoring detects subtle security threats, such as insider threats, privilege abuse, and stealthy malware operations.

Anomaly detection is one of the most essential sub-tasks of AI systems used in cybersecurity, in which AI can determine users' profiles employing ML and identify deviations from the norm. Unlike traditional methods of security where a rule is set and enforced statically, anomaly detection is more adaptive to new threats, and hence is effective in detecting zero-day attacks and insider threats. Adaptive AI models use continuous learning frameworks to remain relevant against evolving threats.

AI refines its threat detection capabilities by integrating new attack data Unsupervised learning is useful to identify new patterns that have previously not been encountered. These deep learning models detect subtle anomalies in network traffic and file system behaviours by identifying outliers in a data set. Adaptive AI models use continuous learning frameworks to remain relevant against evolving threats and refine their threat detection capabilities by integrating new attack data.

User and Entity Behaviour Analytics (UEBA) is an AI-powered approach that detects anomalies in user activity and system processes. For example, if a user suddenly logs in from multiple locations or an unusual device, AI flags the activity as suspicious, or if an employee who normally accesses marketing materials suddenly downloads sensitive financial records, it could indicate an insider threat. By combining UEBA with anomaly detection, AI security models can differentiate between normal user behaviour and potential cyberattacks. To ensure maximum security effectiveness, organizations must implement a combination of AI-based automation, real-time alerting, and human oversight.

AI-powered security systems should not only detect threats but also initiate automated incident response actions to contain potential security breaches. If AI detects malware, it can automatically isolate affected devices to prevent network-wide infections. AI follows predefined incident response procedures to mitigate threats with minimal human intervention. However, human oversight remains critical to validate AI-driven responses and prevent unnecessary disruptions.

Despite its advantages, AI-powered anomaly detection has challenges that must be addressed. AI security models may incorrectly flag normal behaviour as suspicious, leading to unnecessary investigations, and conversely, may miss advanced threats that do not fit predefined behavioural patterns.

4.8 Conclusion

AI automation speeds up the efficiency in the security operations, however, human supervision is still needed to make sure, that AI driven security decisions are correct, ethical and context dependent. Though AI models are extremely good at pattern recognition, they simply don't possess human intuition and contextual awareness. Security analysts should review AI-generated alerts and verify critical threat decisions before automatic enforcement actions are taken. AI security tools should provide transparent justifications for their decisions, enabling cybersecurity teams to validate AI outputs. While AI can detect threats quickly, human analysts should oversee remediation actions, ensuring that AI-driven responses do not cause operational disruptions.

Organizations implementing AI have to balance between AI automation and human expertise in the quest of reducing risk and increasing trust of AI security deci-

sions. The mitigation of AI risks in network security should be done in a holistic manner at the technical, privacy and supervision fronts. It is important to dovetail human oversight with\ the AI security process, a strategy which as a process is aligned with both broader cybersecurity strategies and the adoption of AI driven decisions. Trustworthy, resilient and powerful AI based cybersecurity systems can be built with proactively managed AI security.

The AI driven network security will continue to include a well-integrated risk management process of automated response mechanisms and better threat detection. Prevention, which limits and stops cyber-attacks in real time, will become more and more of a focal point. Today's security model for many of the AI systems is a 'black box'; security teams don't know how the system got to the conclusion it reached. The goal is to help understand how AI threat detections occurs; and thus make AI security models interpretable so that the reasoning behind the threat detection is clear. Perhaps the most important ways to explain AI decisions and to strengthen AI security solutions will be SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). XAI will play a role in assisting what needs to be embraced by AI cybersecurity solutions from a regulatory body's onset in terms of transparency.

As with other critical trends in cybersecurity, we find there is another critical trend that is coming, that of Zero Trust Architectures or Zero Trust AI (ZTA). Traditional security models based on the trustworthiness of the running entities in the network need enhancing to protect from insider threats, credential theft, and incidents of horizontal movement. In contrast, the 'never trust, always verify' principle based zero trust security frameworks are based on granting network resources only if users are authenticated and verified. When it comes to zero trust frameworks, AI plays an important role in the automation of identity authentication, behaviour-based authentication, and real time access control decisions.

The leap forward in automated cybersecurity response is through networks which evolve using their self-healing properties. The autonomous threat mitigation or AI powered autonomous threat mitigation, that enables security systems to heal themselves from cyber-attacks without human intervention, powers the source of self-healing networks. For instance, the intruder could start by creating malware that infects the endpoints, and AI automatically isolates the affected endpoints, retracts the malicious traffic as well as triggers system recovery processes to make downtimes minimal and less damaging. The predictive analytics through real time threat intelligence and past attack patterns will be a part of self-healing network where the security personnel would be given the ability to predict cyber threats with the use of AI. But as the attacks become more frequent and complex, these self-healing networks will become more imperative because human security teams would be increasingly burdened doing this.

Traditional AI security models rely on big datasets, that can be privacy and compliance problems. In federated learning, AI models are trained on aggregate data

from various decentralised devices without exposure of sensitive data. AI based deception technologies are used to draw the attackers into the controlled environment for analysis uncovering new attack vectors. Governments and industry regulators are setting the standards on the making of AI driven cybersecurity solutions, and these AI driven cybersecurity solutions have to be made in a way that is fair and transparent. Future AI security models adhering to global regulations like GDPR, and soon out to be released AI governance frameworks, will ensure a balance between security innovation and privacy protection. Finally, XAI, zero trust architectures, self-healing network and AI driven deception technologies will influence the future of AI network security. The outcome of this will be more transparent, more adaptive and more autonomous AI security systems that improve organisations' capacity to detect and eliminate cyber threats.

As the application of AI into network security continues to grow, there are great advantages but also simultaneously numerous risks. AI can accelerate threat discovery in the network, find automated responses, and enable predictive analysis, with which organizations can prevent cyber incidents from manifesting. However, there are certainly risks also – adversarial attacks, biased decisions, black box issues and automation failures. Organizations have to be careful in the trade-off of having a well-balanced AI security strategy that does not make safeguards weaker but makes them stronger. AI security can be adopted in a strategic and risk aware manner, to be used beneficially as a force multiplier, and not becoming an attack vector.

References

Abdi, A. H., Yaseen, O. M., Rehan, H., & Ahmad, M. (2024). Security control and data planes of SDN: A comprehensive review of traditional, AI, and MTD approaches to security solutions. *IEEE Access: Practical Innovations, Open Solutions*, *12*, 69941–69980. https://doi.org/10.1109/access.2024.3393548

Abed, A. K., & Anupam, A. (2023). Review of security issues in Internet of Things and artificial intelligence-driven solutions. *Security and Privacy, 6*(3). https://doi.org/10.1002/spy2.285

Abuali, K. M., Nissirat, L., & Al-Samawi, A. (2023). Advancing network security with AI: SVM-based deep learning for intrusion detection. *Sensors*, *23*(21), 8959. https://doi.org/10.3390/s23218959

Ait, N., Rahman, A., & Shah, M. (2020). A secure AI-driven architecture for automated insurance systems: Fraud detection and risk measurement. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 58546–58558. https://doi.org/10.1109/access.2020.2983300

Bellamkonda, S. (2024). AI-driven threat intelligence for real-time network security optimization. *International Journal of Computer Engineering and Technology (IJCET)*, *15*(6), 522–534. https://lib-index.com/index.php/IJCET/article/view/IJCET_15_06_044

Benzaid, C., & Taleb, T. (2020). AI for beyond 5G networks: A cyber-security defense or offense enabler? *IEEE Network*, *34*(6), 140–147. https://doi.org/10.1109/mnet.011.2000088

Bobde, Y., Paliwal, P., Shah, A., & Kulkarni, S. (2024). Enhancing Industrial IoT network security through blockchain integration. *Electronics*, *13*(4), 687. https://doi.org/10.3390/electronics13040687

Chattopadhyay, R. (n.d.). *AI-driven adaptive encryption: Transforming financial data security in the age of digital banking*. IRJAES. http://irjaes.com/wp-content/uploads/2024/12/IRJAES-V9N4P208Y24.pdf

Choudhury, N. R., Paul, S., & Ghosh, S. (2024). Comparative analysis of traditional vs. AI-driven network security. In *Advances in Wireless Technologies and Telecommunication* (pp. 107–128). IGI Global.

Illiashenko, O., Gerasin, V., Borovkov, A., & Morozov, V. (2023). Security-informed safety analysis of autonomous transport systems considering AI-powered cyberattacks and protection. *Entropy*, *25*(8), 1123. https://doi.org/10.3390/e25081123

Kaul, D. (2021). AI-driven decentralized authentication system using homomorphic encryption. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, *12*(2), 74–84. https://lib-index.com/index.php/IJARET/article/view/IJARET_13_03_008

Kannan, Y. (2024). AI and machine learning for network security: Applications and case studies. *International Journal of Artificial Intelligence & Machine Learning (IJAIML)*, *3*(2), 1–13. https://lib-index.com/index.php/IJAIML/article/view/IJAIML_03_02_001

Krti, T. (2025). CyberSentinel: An emergent threat detection system for AI security. *arXiv preprint*. http://arxiv.org/abs/2502.14966

Li, G., Zheng, P., Wu, D., & Yang, F. (2025). AI-driven solutions for proactive network security and threat detection. *IET Conference Proceedings*, *2024*(21), 756–760. https://doi.org/10.1049/icp.2024.4311

Nay, T. (2024). Enhancing IoT security with AI-driven hybrid machine learning and neural network-based intrusion detection system. *Babylonian Journal of Artificial Intelligence*, 158–167. https://doi.org/10.58496/bjai/2024/017

Patel, R., Sharma, K., & Kumar, M. (2024). Zero Trust Security Architecture raises the future paradigm in information systems. *Informatica and Digital Insight Journal*, *1*(1), 24–34. https://firstcierapublisher.com/index.php/Informatica/article/view/77

Dasaratha Rao, D. A., Waoo, A. A., Singh, M. P., Pareek, P. K., Kamal, S., & Pandit, S. V. (2024). Strategizing IoT network layer security through advanced intrusion detection systems and AI-driven threat analysis. *Journal of Intelligent Systems and Internet of Things*, *12*(2), 195–207. https://doi.org/10.54216/jisiot.120215

Ranganathan, C. S., Karthikeyan, R., & Ramachandran, V. (2023). Network security in cyberspace using machine learning techniques. In *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1755–1759). IEEE.

Rangaraju, S. (2023). Secure by intelligence: Enhancing products with AI-driven security measures. *EPH – International Journal of Science and Engineering*, *9*(3), 36–41. https://doi.org/10.53555/ephijse.v9i3.212

Rashid, M. M., & Yaseen, O. M. (2025). AI-driven cybersecurity measures for hybrid cloud environments: A framework for multi-cloud security management. *International Journal on Engineering Artificial Intelligence Management, Decision Support, and Policies*, *2*(1), 30–39. https://submissions.adroidjournals.com/index.php/ijaimd/article/view/39

Rehan, H. (2024). AI-driven cloud security: The future of safeguarding sensitive data in the digital age. *Journal of Artificial Intelligence General Science (JAIGS)*, *1*(1), 132–151. https://doi.org/10.60087/jaigs.v1i1.89

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero Trust Architecture* (NIST Special Publication 800–207). National Institute of Standards and Technology.

Salem, A. H., Bhuyan, M. H., Ahmad, I., & Nanda, P. (2024). Advancing cybersecurity: A comprehensive review of AI-driven detection techniques. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00957-y

Sarker, I. H. (2024). *AI-driven cybersecurity and threat intelligence*. Springer Nature.

Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-driven cybersecurity: An overview, security intelligence modeling and research directions. *SN Computer Science*, *2*(3). https://doi.org/10.1007/s42979-021-00557-0

Siddique, M. R., Sharma, D., & Sahu, P. K. (2024). Integrating machine learning-powered smart agents into cyber honeypots: Enhancing security frameworks. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)* (pp. 1–7). IEEE. http://dx.doi.org/10.1109/I2CT61223.2024.10543399

Suram, K. K. (2024). Improving cloud security via AI-driven cloud infrastructure automation and human-AI collaboration in IAM. *International Journal of Computer Engineering & Technology*, *15*(6), 1734–1742. https://doi.org/10.34218/ijcet_15_06_148

Umoga, U. J., Etim, S. E., & Chukwuemeka, M. (2024). Exploring the potential of AI-driven optimization in enhancing network performance and efficiency. *Magna Scientia Advanced Research and Reviews*, *10*(1), 368–378. https://doi.org/10.30574/msarr.2024.10.1.0028

Yu, Q., Xu, H., Wang, H., & Wang, M. (2020). An immunology-inspired network security architecture. *IEEE Wireless Communications*, *27*(5), 168–173. https://doi.org/10.1109/mwc.001.2000046

Zhang, J. (2021). Distributed network security framework of energy internet based on internet of things. *Sustainable Energy Technologies and Assessments*, *44*, 101051. https://doi.org/10.1016/j.seta.2021.101051

Zhao, J., Masood, R., & Seneviratne, S. (2021). A review of computer vision methods in network security. *IEEE Communications Surveys & Tutorials*, *23*(3), 1838–1878. https://doi.org/10.1109/comst.2021.3086475

Gregory Bramble

5 How to architect artificial intelligence systems that put the human in the loop and don't overpromise

5.1 Introduction

When it comes to balancing risk and reward in Artificial Intelligence (AI), there is a long list of relevant topics and much research that successfully boosts accuracy and confidence scores, analyzes trustworthiness, identifies and mitigate bias, and establishes appropriate security and governance policies. The research in these areas is truly astounding, and AI is developing at a breakneck pace because of the impressive advancements being made. With all these advancements, engineers in AI research are in a position, now more than ever, to build systems that humans can trust and use to make their jobs and lives easier. To go even further beyond the current state of AI and build truly useful and innovative AI systems of the future, AI engineers must make use of all these innovations, while keeping in mind that AI itself will never be perfect.

The main reason that AI will never be perfect is that Artificial Intelligence implicitly proposes to be able to answer any question or any problem that can be posed at any time, just like human intelligence. To show how difficult it would be to achieve perfect accuracy on such a task, we could restrict the domain of questions posed to the AI to simple ones like "What is 2+2?", or factual ones like "What are the three primary colors?" On these two questions, there are many Artificial Intelligence systems that will answer "Four.", and "Red, blue, yellow.", and thereby achieve perfect accuracy within this controlled domain. However, even on these questions, there are some Artificial Intelligence systems that will answer incorrectly, including among the large language models of today. It is possible that one of them could answer incorrectly for various reasons, including hallucination, a poor training set, a bug, a missed corner case during human tuning, and other possibilities. When we then remove the restriction, and, for example, allow the input to be any question, even still factual ones, we may ask a question with very little available public information about it, and on which the model may not have been trained; a question to which very few people know the answer. Often, even the best models of today will attempt to answer and may answer confidently, but incorrectly. Going even further, we have questions to which even humans cannot confidently agree on an answer; questions about what will happen to the stock market, or what will happen after we die, or other questions about the future. A very intelligent entity might answer some of these questions more correctly than one that is less intelligent, but a perfect entity could answer them all correctly. Given the reality that we are speaking about, and the fact that Artificial In-

https://doi.org/10.1515/9783111595306-005

telligence is created by imperfect beings who are never 100% correct, and that all of the information that the Artificial Intelligence is trained on is imperfect, it seems impossible and would require an impossible stroke of luck for any Artificial Intelligence system to answer such questions and in so doing, be perfect.

Therefore, taking it as a reasonable assumption that we cannot ever create perfect Artificial Intelligence, we are presented with a more palatable question, which is itself an architectural and engineering question: how can we create Artificial Intelligence systems that allow humans to interact and understand, and even allow the system to understand the human, so that the humans benefit from the Artificial Intelligence while mitigating the risks of the system's imperfections? This is a difficult task, and discovering possible pathways to answer will be the subject of this chapter. The good news is that we already have a lot of experience in the research community around human interaction with computers, both with and separately from Artificial Intelligence, to rely on. Humans come with a diverse array of experiences, education, and understanding, and the system needs to be built in such a way as to inform the human of what it is doing and how it operates, in order that the human can interact with the system and use its results. This is not only an AI question, but some of the promise of AI makes this question more difficult because many humans may come to an Artificial Intelligence system expecting perfection. All of the advancements in AI, and especially in research that mitigates some of the imperfections of AI, including metrics around trust, governance, confidence, and improving accuracy, must be displayed to the user in a digestible and understandable way to help them see when a model is performing well and when it is not. How this looks to the user evolves as technology advances, as human understanding of AI advances, and as architectures adapt to these two moving frontiers of technological advancement and growing human understanding of new technology. The best AI architectures build effective bridges across these two frontiers. This chapter proposes design and engineering principles to build trustworthy, human-centered AI systems that acknowledge their limitations and promote responsible use.

5.2 The challenge of measuring human-AI understanding

In the context of managing risk of AI systems, bridging the gap between technological advancement and growing human understanding is hard to measure. The problem with measuring whether or not the system will perform well prior to launch is difficult because most of the ways to tell are heuristic and very human-specific. A lot of times, we do not know that we got the bridge perfectly right until the application is in production and users either like it or they do not. If you look at the recent example of OpenAI's ChatGPT, they seem to have gotten one part of the balance right as they ex-

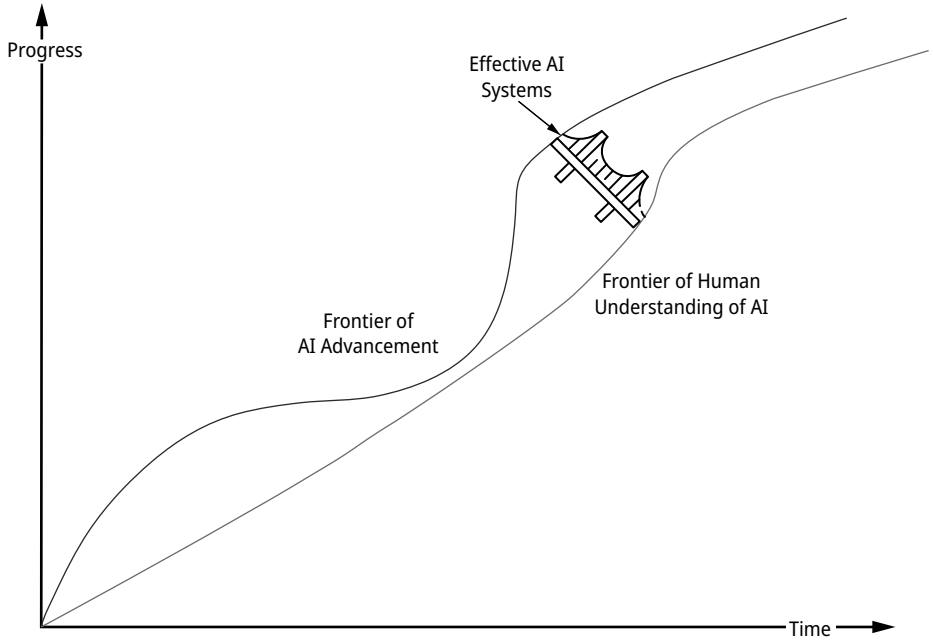


Figure 5.1: Building Effective AI Systems by Bridging Expanding Frontiers.

posed a very large advancement in technology along with an interface to enhance human understanding of what was going on. This chat interface proved to be intuitive to users and gave them the ability to communicate back and forth to clarify questions. This interface combined with the ground-breaking technology to appeal to users and achieve acceptance and adoption. In spite of their success, ChatGPT had a serious roll-out issue, and one that remains: hallucination. With hallucination, a large language model says something that is factually untrue but says it as if it is true. Basically, just happily going along telling falsehoods. Despite that, ChatGPT did provide so much good knowledge that users seemed able to forgive this massive shortcoming, either by checking the answers themselves separately or using their own judgement to determine the correctness. According to OpenAI usage reports, ChatGPT maintained rapid user growth despite known hallucination issues. Comparing the Large Language Model (LLM) to humans that also make mistakes, users seemed to be forgiving. This may also be partly because ChatGPT's initial free offering may have helped build goodwill and set appropriate user expectations. In addition, the user experience in ChatGPT, especially around its ability to respond to user complaints and answer clarifying questions were a great help to human understanding.

"The basic idea behind the Transformer is to use self-attention to encode the input sequence and produce a sequence of hidden representations, which can then be decoded into an output sequence. Self-attention allows the model to attend to differ-

ent parts of the input sequence at different levels of abstraction, which helps it capture long-range dependencies and relationships between different parts of the sequence." (Ray, 2023)

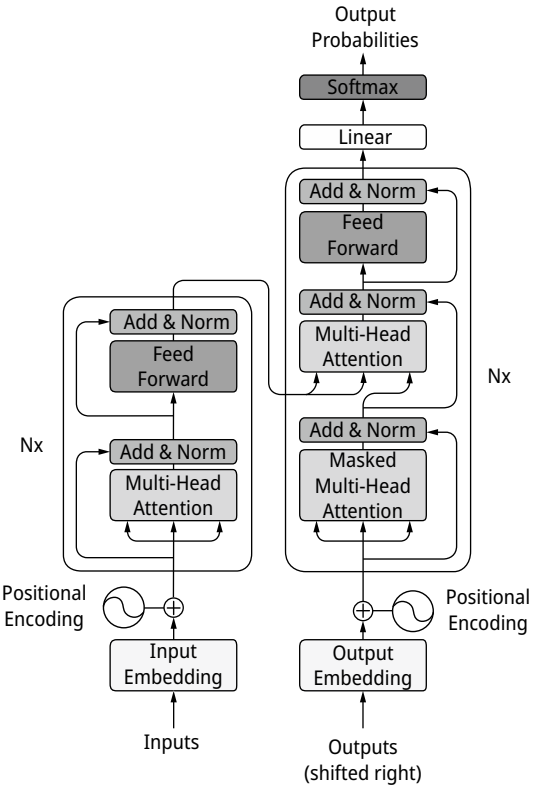


Figure 5.2: The Transformer Model Architecture from 'Attention is All You Need' (Vaswani, 2017).

It is this self-attention innovation that led to the massive success of ChatGPT. It is interesting to consider that prior to this innovation, the model was equally if not more intelligent than after; meaning it had all the same information stored inside of it before and after. What really changed in this innovation was that the model was better able to focus on what the human user was saying and respond to the user better. So, one could argue that the explosion of AI that came from ChatGPT was not really because the model all of the sudden got smarter, but rather, because it effectively aligned user experience with underlying model capabilities. Taking this argument a bit further would also tamp down some of the fears around the rise of Artificial General Intelligence, which is anyway just a roughly defined concept and not something that is likely to come about in any of the ways currently being discussed. Raw intelligence is not actually something that we are in short supply of, but rather it is finding

that sweet spot of combining knowledge, information, and good AI modeling with human understanding that always brings us further in terms of technological advancement in AI.

5.3 Background: What is AI?

Before we delve deeper into trying to discover the best ways to bridge the two moving frontiers in AI advancement, to better ground the discussion, we must define Artificial Intelligence. For a moment, try to think of your own definition. It is not so easy to define. We know from computational theory that machines can solve two types of problems, those that are exponential (Non-Polynomial or NP) in time, versus those that are polynomial (P) in time. All NP-hard problems are not solvable by classical computers. So, in some sense, Artificial Intelligence is any attempt to use heuristics, models, or other means to solve NP-hard problems using a classical computer (Starner, 2018). Looking a bit more historically, Alan Turing in 1950 proposed 'The Imitation Game' as a way to define Artificial Intelligence. In this game, sometimes referred to as the Turing Test, the setup is that he would have an interrogator in a different room from a person and a machine and see if the interrogator can tell them apart. Turing wrote in 1950: "I believe that in about fifty years' time it will be possible to program computers ... to make them play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning." In Turing's world, if a computer can successfully play the Imitation Game, then we have AI. Taking these definitions into account, we can get a good sense of what people mean by AI. Perhaps more important than getting the definition right, we need to try to get the bridge between the two moving frontiers of technological advancement and human understanding right in order to build great AI systems.

It is helpful to look more into the history of AI to see that its meaning has changed and evolved for people over time. At one point, we thought of Automatic Navigation as AI. The first Global Positioning System (GPS) based maps applications wowed users and changed the way we get around the world. Also, when spell check was first released, it was not thought of as just a red squiggly line under your words in your favorite word processor, it was a game-changing AI feature. Also, with spam filtering, which we take for granted today, at the time it was a major differentiating AI offering that made a big difference in the usability of email. DeepBlue was an IBM AI system that beat chess champions, with algorithms better than brute force search. The world tuned in and was amazed to see the computer competing at that level using AI. More recently, Machine Learning became more and more important to AI. Technologies began to make AI easier to use. Watson beat Jeopardy champions, and that win marked an important moment in the rise of AI, showing that AI systems could effec-

tively and intelligently communicate with humans using natural language. And most recently, with the rise of large language models, we saw the ability for human users to be able to directly converse with AI models in applications like ChatGPT.

At this point, let's take a step back and think about what AI really is throughout history. Certainly, spam filtering would not be considered as state-of-the-art AI today, but at one point in time, it was. So, in some sense, the things that make up the category of 'What is AI?' really do change over time. Even though it might be difficult to believe at this moment in time, the large language models of today also one day may not be considered on the frontier of AI, or even really AI anymore as we will one day think about it. So then, how is AI really defined? As I have said, it is really in the architectures and systems that build the bridge from the latest technologies to the human understanding of them. In this light, we can think of AI as having moments throughout history where we really saw it shine, in those big successes like the Watson Jeopardy win and the release of ChatGPT and looking forward we can try to find where that next big and important bridge between technological advancement and human understanding will be built. As AI architects and engineers, that is our job.

Yet, I am not satisfied to conclude that we are only in just another moment in the history of AI. Although in some sense that is the case, there is a certain different feeling about the magnitude of the current AI moment. Why now? Suddenly, we have had a convergence of multiple factors that has led to a fast advancement in AI technology in recent years. The first is the availability of more and more powerful Graphical Processing Units (GPU's) that large businesses can buy. GPUs essentially were built to draw graphics faster in machines, and therefore they do certain complex mathematical operations much faster than classical CPUs. Companies in the business of building chips for screens and video game consoles therefore were in the sweet spot to start selling their technology to AI companies, and this is a reason for part of the magnitude of the current moment. Along with that, the cost was low enough for GPUs, while definitely not cheap, for large technology companies that have a lot of cash to build huge AI data centers full of GPU's, to train and build massive AI models. Another thing maybe not thought of as much was that we had the internet just sitting around, which is essentially almost forty years of human cataloging of all the data and information known throughout history. That endeavor was part of the information age's desire to learn more and share more about life itself, but it had the unintended effect of creating exactly the kind of dataset that massive AI models need to reach a new level of both technological power, and the ability to converse with human users. Whether or not that is truly the intended use of the internet is another question, and the governments of the world need to sort out through legislation and litigation whether or not it is a fair use of the internet. This is another aspect of the riskiness of deploying AI software: do the builders have the right to use the underlying data? We will touch more on this difficult subject later. The final piece of the magnitude of the current AI moment is made from marvelous engineering in companies that built these massive AI models and made them work through advanced engineering pro-

cesses that both harnessed the huge volumes of data on the internet and the human inputs of engineers and other contributors.

All of these things enabled the technology we have in the current AI world to get to human users effectively. The large companies are able to offer these models that are expensive and difficult to train. As a sidenote, there is a recent advancement from DeepSeek indicating it may be possible to train models more affordably and possibly more democratically. However, even absent that innovation, in the current moment, small companies can still be AI companies by using the large language models hosted by model providers for a small and affordable startup cost.

If we look deeper into what is so compelling about the large language models of today, we see that LLMs like ChatGPT can pass all of the major tests, SAT, Law, and beyond (OpenAI, 2023). This itself is astounding compared to the systems of just a few years ago. The power of these models comes in information synthesis from many sources. One way to think of the operation of these models is that the LLM finds the likeliest path through an alphabet soup of text. This is not magic, but this is what it does so well. One character or short string of characters at a time, the large language model predicts what comes next and takes into account very much what was said recently. By doing this, it is able to find patterns and generate language that is both fluent and often accurate, although it does still lag human expertise in many domains. By passing those major exams it showed its ability to achieve above average human intelligence, but not necessarily expert level human intelligence in all domains.

Therefore, our challenge remains how to connect this technology to human users effectively. One major pitfall, as we mentioned in the introduction, is that the GPT models sometimes hallucinate without informing the user. This puts sometimes a large unknown in the midst of what seems a fluent and intelligent conversation; leaving the user asking him or herself, "Is this correct?" The best AI systems of tomorrow, as today, must make sure that the human is in charge of verifying correctness as easily as possible in a way that is helpful to human endeavors. Another challenge that is arising is that humans should be careful not to use these AI models too much, because it is not such a great idea to outsource your intelligence. Human intelligence develops through rigorous engagement with challenging thoughts and ideas, and so as a personal matter, it is best that we remain engaged and not outsource our key educational and intellectual pursuits to AI.

We mentioned earlier that there is another important risk to consider around whether or not AI companies have the right to use the data underlying their models. Especially with LLMs just trained on the internet, the companies creating them are copying and reusing wide swaths of copyrighted material to do so. What is the ethical way to handle citations and copyright issues? Perhaps there is a legal fair use of these AI models, and seemingly there must be new paradigms created where the copyright holders are fairly compensated for the value they contribute to these models.

Another emerging area of risk, and one pointed out by SEC Chair Gary Gensler is: How do we regulate across companies if they all share one or few models? Think of a

case where we have AI applications all helping with investment decisions, and they are all using the best AI model of today to inform their decisions. As we mentioned, even if that model is the best, it cannot be perfect, and therefore it has some or many blind spots. If many companies are all using that same model underneath, then all companies would have the same blind spots. And, therefore, when reality hits and something happens that all of the companies missed, we could have more severe negative effects that we would not have had in a world where all of the companies were making decisions more independently (Bracken, 2024). One mitigation for this is the recent trend around open-source AI. It does seem critical that the democratization of building AI occurs as it will help us allow companies to differentiate their models based on their particular expertise, and therefore not only avoid bad situations like the ones mentioned, but also to improve the quality of AI.

So indeed, we do have the technology, in fact we have something close to what Alan Turing was dreaming about. With that in our hands, it seems that we have not yet scratched the surface of all of the potential applications. The first generation of LLM powered applications seems to be something like a very intelligent personal assistant that can understand my thinking and answer me at will. That is good, but the current challenge then is for us to come up with the next generation of great AI applications. That brings us back to the main topic of this chapter, and raises questions like 'What makes a great AI product?' Again, the AI enterprise has to find a good balance. The right level of interaction is critical, across the frontend and backend, and between AI and the human. The opportunity – and the responsibility – to shape new ways of doing business, and even new ways of making a living is upon us. We must embrace this now more than ever. We must look positively at the technology and look at building the bridge between the technological frontier and the frontier of human understanding.

5.4 Practical experience bringing AI systems from idea to reality

Let us think of the world of research as the frontier of the ever-advancing technological landscape. We can measure how many publications a group is producing at top conferences, and how many patents a group is publishing, to get an idea of how they are faring in their technological advancement. Let us then roughly think of the product group as the frontier where human understanding of technological advancements subsists. We can measure how well a product group is faring at turning future technology into human understandable products by market share and revenue. In fact, a good standard for the quality of a technology enterprise is whether or not the enterprise is having state-of-the-art publications turn into state-of-the-art products. But how does an enterprise invent and create the next generation of technology? The

product groups often may not be aligned to the research groups; but may be a year or a few years behind. This is normal; research, by definition, is ahead of practice. But this becomes one of the key challenges in building this bridge across the two frontiers: while the product group is advancing the human real-world understanding of the technology, they do not know the new technology yet at all. In fact, if research is really doing its job, then they are dreaming up the technology of the future. A big responsibility therefore falls on research to not only provide next generation publications and technology, but to provide the vision for the product group to execute in the future. This is a significant challenge, and one that sits at the heart of technological enterprise and also at the center of the design of human in the loop systems for technology in general and for AI in particular.

In my experience, initially research groups' ideas often have little to no advocacy in the product groups. This is really because the product teams have limited awareness of emerging technologies, which may not yet be formalized or communicated. It is therefore research's responsibility to advocate for their ideas. Initially, they may be the only ones who are working with a given new technology as their primary effort. Product counterparts are always busy doing their day jobs; still helping but aligned to an existing product. Since the research team therefore is the only team 'wearing the colors' of the new idea, the responsibility falls on them in design meetings, in architecture meetings, and with product management to teach the new technology and co-create the product. This is all about providing product vision, which is a different skill than pure research and innovating. It requires spending the time to develop the knowledge of the existing products and offerings, and why things are the way they are. It requires knowing the product is how it is for good reasons, and that while research may have a cutting-edge system, the product team may not understand it yet at all. It is about becoming the authority on the topic that can provide thought leadership and direction. Therefore, three good criteria for a research-minded organization to keep in mind to stay on track are:

1. Am I publishing or helping the team publish at top conferences?
2. Am I knowledgeable on current offerings in my field of expertise?
3. Am I developing a vision of what I think that offering set should look like in a year, and in several years, and where my research fits in?

5.5 Engineering principles that enable risk mitigation

Assuming all goes well, and we are able to effectively get our idea from inception to productization, we must keep in mind that mitigating risk starts by thinking of the human from an idea's very first conception all the way until its final deployment in production. The practice of making technology work well for humans is engineering,

and therefore the research organization that is in the ideation business ought to practice solid and simple engineering principles from the very beginning. In fact, although it is impossible to measure because it is so pervasive, good engineering is perhaps the best risk mitigation strategy we have. Research has shown that a bug in a requirement costs ten times the amount to fix as a bug in the code itself, which if we extrapolate to also cover a bug in the perception of inherent risk in a given deployed system could be even more expensive. So, doing great engineering becomes that much more critical. That starts with getting the foundation right. To do that, the systems should be easy to install, easy to understand, and easy to maintain. To achieve those ends, the systems should have Application Programming Interfaces (API's) based in key concepts from publications. By looking to publications for key common concepts, engineers can make API's that are useful for everyone.

In addition, it is important for AI systems from their conception to rely on open-source standards. One of the great things about open-source software beyond the free exchange of engineering is the convergence on open standards. We tend to take this for granted, but for me personally I learned from a young age about the importance of standards when I learned about the Great Baltimore Fire of 1904. This fire was a massive disaster that caused the burning of huge parts of Baltimore City. In the early stages of the fire, help came from many surrounding cities including massive trucks from Annapolis, Harrisburg, Wilmington, Philadelphia, and New York. It seemed that all of this additional help would put a quick end to the disaster. But when all of the trucks tried to attack their hoses to the local Baltimore fire hydrants, none of them fit. If only a handful of these nozzles had fit the hydrants, the entire fire would have been doused and put out almost immediately. There was no shortage of water in the hydrants, nor was there shortage of trucks, hoses, and firefighters. The only thing in short supply was good engineering and common standards. Because of this, over 1,500 buildings, all power, telephone, and electrical facilities in an area of over 140 acres in the business area of Baltimore City were destroyed before the fire finally burned out. It took 30 hours for the blaze to go out on its own. This tragedy reminds us of one key engineering principle, which is also a risk-mitigating strategy, that it is critical to have interchangeable parts based on open standards. And maybe you think that AI software engineering is disconnected from fire hydrants, but in AI systems, it is indeed quite important. The Great Baltimore Fire of 1904 exemplifies the dangers of lacking interoperability; a principle that also applies to incompatible AI components and standards today. For example, a given AI class might implement the open source scikit-learn estimator API, which allows nesting and easy utilization of existing libraries. This gives these AI systems powerful cross validation, certain metrics, and folding strategies for free. Whereas if the initial engineering team ignored these open-source standards, they might have to spend quite a lot more time, energy, and maintenance to achieve the same functionalities, if they ever can at all.

Beyond getting the foundations right in engineering, one of the major drawbacks of more complex Machine Learning and AI systems like LLMs, Support Vector Ma-

chines (SVM's), Random Forests, and others is that since their modeling approach is so complex, it is difficult for a user to understand the rationale for any given prediction, which is a major barrier to human confidence in such machine intelligence. This is a problem in the machine learning and AI community, which I believe will only worsen, as models get more and more complex in the quest for better, more accurate, predictions. What approach can we take when the purported experts in these models admit they themselves do not fully understand them?

5.6 Importance of benchmarking

Benchmarking is about convincing ourselves, our researchers, and the product team that our technology works. It is going beyond the old developer adage: 'It works on my machine.' There is that moment in the life of the developer when they have very bleeding edge code, and people clamoring for results. Under that pressure, they hesitate to learn a new framework and add benchmarking code that could cause more bugs. A key for any successful benchmarking framework is to get the developer to simply hook in the framework during that moment. and then hopefully that same hook up can be reused later for the cluster runs. The framework should provide datasets already hosted, possibly a decorator to do runtime and memory for starters, and it should 'just work' locally. Once this is setup, a lot of good things happen. First, the developer gains confidence in his or her code. Second, the researcher gains confidence in his or her idea or modifies it to make it even stronger. Third, management begins to take note, and the real-world impact begins with productization discussions, followed by marketing, deployment, and hopefully successful usage. Perhaps most importantly, if the team comes up with truly good benchmarks which are evaluated on real-world ground truth, they are then able to deliver a system that truly says "yes" when it means "yes," and "no" when it means "no."

Another key principle of good software engineering to mitigate risks inherent in complex systems is: Keep it simple. Keeping it simple is often a great challenge – so try to spend the time to try to keep it simple, of course, within reason and when possible, and extend where needed. It is not always possible to stick to the existing conventions, and in those cases, it is good to reuse what exists and extend. This is especially important due to another key limitation of LLMs that we have not mentioned, which is limits on the amount of input text allowed. As we mentioned, LLMs are effective because they pay a lot of attention to recent text. This also brings an issue because there is a limit to the amount of text that they can pay this much attention to. Therefore, for expert-level applications, the LLM may not be able to attend to enough of the pertinent information to make an expert-level decision simply because it is not possible to give that much information to the LLM. Innovation in LLM design will continue

to make progress on this front, but ultimately it will also be up to good engineering of the systems to help bridge this gap.

5.7 Ethical implications of deploying AI systems

Let's take another area where risk management is going to be very relevant for AI: the ethical implications for autonomous systems. While ethical principles are universal, their application varies across legal and cultural contexts. Democratic societies in the world today, may and do have differing public opinions on what is good, and therefore different laws governing ethics. This means that no one country is truly and completely ethical. As a governing principle, it seems expedient to try to leave ethical decisions in the hands of individuals as much as possible, and only to use correction in cases where the humans are unethical and break laws. With AI however, this principle seems to not hold up anymore, because the humans who create the AI systems are forced to answer ethical questions before the individual human users are even faced with the ethical questions. Being placed in such a position to answer ethical questions days or even years before a life-or-death situation occurs seems an almost superhuman power. How to use this power effectively, and hopefully, to instead decentralize this power into the hands of the users as much as possible, seems the best course forward. But there is certainly not a clear line to draw here, and I do not propose to be able to definitively draw it.

5.8 Design as the intersection of research and product, and humans and machines

Building AI capabilities across an organization can often lead to fragmented solutions spread across multiple locations, with little to no version history, traceability, or lineage, and minimal to no collaboration among key stakeholders throughout the process. In addition, different teams with differing AI proficiencies often may not work with other teams because they view them as inferior or have no available platform for quality collaboration. Building quality AI also requires that all stakeholders are constantly learning in their daily working environment so that they are becoming more proficient in understanding the key principles of AI systems and how to build better AI pipelines. Our future AI systems therefore must enable key stakeholders to build AI with these very things as fundamental aspects of the design. Focus must be placed on technologies that dramatically condense the AI lifecycle by enabling data scientists and AI developers to focus on only the tasks that require human expertise, while the system handles the rest of the tasks. Typically, the non-automated data science pro-

cess takes human data scientists weeks or months to build models and digest insights from data. We want to focus on systems that can take days or less because data scientists can directly use top performing pipelines, study results to gain ideas, and improve the AI pipelines easily.

One key element of the success of our AI systems must be that we do not endeavor to replace human workers, but rather to collaborate with humans. Especially with the recent interest in Agentic AI, where we build AI systems that begin to act autonomously in the world, we must endeavor to build and expand upon our current collaboration capabilities, both among humans, and between human and machine. The collaboration capabilities launched in today's AI systems are among their most important features. In Agentic AI, where we give autonomous decision-making capabilities to the machine through LLM-powered agents, getting the balance right seems more important than ever. It was one thing to make sure that the human in the loop understood what the system was doing when it was writing creative songs, suggesting more concise code, or correcting grammar in an essay, but how much more important will that be when the system is trading stocks, purchasing airline tickets, or buying a new car? Can we ever trust a system to make such decisions on our behalf? The answer is not a simple binary but truly depends on how well we get all relevant parties into the room when we are designing our applications of the future, and how well those applications balance the two competing frontiers we have been discussing in this chapter.

Building effective, risk-mitigated AI systems requires focusing on all of the parties and processes that lead to these AI systems. It turns out, the major players, as we have seen, are research and product, and humans and machines. One place where all of these come together in a successful AI enterprise is in Design. There is a classic design meeting called a Four-in-a-box, which combines User Interface Development, User Experience Design, Product Management, and Backend Development. I like to run a modified Five-in-a-box, where especially in the early sessions, Research has a leadership seat at the table. By the end of the project, Research typically does not only no longer need to lead, but even no longer needs to attend, and should be off ideating for the next great technology. But in the beginning, their role is critical, as they basically need to teach the new technology to all of the groups, starting with the User Experience Designers, because none of them can truly understand it especially if it is groundbreaking.

Historically, even before AI, applications and application design has been a main driver of good Human-Computer Interaction. Simple API design often leads to widespread adoption and usage. Libraries like PyTorch, and more recently Langchain and CrewAI, achieved widespread adoption not only because of their relevant subject matter but just as importantly because of their simple and clean design patterns that are easy to adopt and use. With the rise of LLMs, more and more, the models themselves are acting as applications, with the design processes needing to be part of the fine tuning and human feedback cycles. Where systems like ChatGPT seem to have gotten

the balance mostly correct, earlier systems like Microsoft's Tay got it terribly wrong (Wolf, 2017). A much earlier conversational agent, Dr. Sbaitso, was far less intelligent than ChatGPT, but was a successful application because it did get the human interaction piece right (Candello, 2016). Sometimes, it is just about saying the right thing when the model knows the answer, and saying it is not sure when it does not.

5.9 Conclusion

We are at a very exciting, and very risky, time in the history of AI. There have been many big moments in AI history, but this one is among the most important because of many factors that have led to very fast advancement in AI technology. We can chat with models that have human-level intelligence that we did not have just a few years ago. Building effective, risk-mitigated AI systems will require great engineering that helps bridge the gap between the fast-moving technology and the human users who demand and need to understand the systems. In this chapter, we explored the reasons for doing this and some of the best ways to go about it. Ultimately, it requires everyone to push for greater quality in research and product leading to great systems that connect machines and humans in new ways that put the human users in the center of the AI and therefore at the center of the world they are pushing forward in the best and most profitable ways. As we've seen, effective AI design requires more than innovation, it demands responsible engineering.

References

Partha Pratim Ray. (2023, April 14). *CHATGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope*. Internet of Things and Cyber-Physical Systems. https://www.sciencedirect.com/science/article/pii/S266734522300024X#bib94

Starner, T. (2018, September 12). CS 6601 Artificial Intelligence.

OpenAI. (2023, March 15). GPT-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf

Bracken, Matt. (2024, January 18). *SEC chair Gensler Sounds Alarm on risks of large AI-fueled financial models*. FedScoop. https://fedscoop.com/sec-gary-gensler-ai-models-financial-sector/

Wolf, M. J., Miller, K., Grodzinsky, F. S. (2017, September 25). *Why we should have seen that coming: Comments on Microsoft's Tay*. ACM SIGCAS Computers and Society. https://dl.acm.org/doi/abs/10.1145/3144592.3144598

Candello, H., Pinhanez, C. (2016, October). *Designing Conversational Interfaces*. https://www.https://www.researchgate.net/publication/333022441_Designing_Conversational_Interfaces

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, İ., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Dron M. Mandhana and Krista N. Engemann

6 A structured approach to evaluating AI capabilities for risk-managed implementation and adoption in organizations

6.1 Introduction

Artificial intelligence (AI)-enabled technologies are increasingly deployed across industries for a multitude of tasks, including data, text, and video analysis, text-to-speech and speech-to-text conversion, predictive decision-making, rule-based automation, and business process optimization (Chowdhury et al., 2023; Faulds & Raju, 2021; Schmid et al., 2021). The rapid uptake of AI tools is driven by the promise that they can produce higher-quality decisions than human experts, improve work efficiency, and enhance both employee and customer engagement (Bughin et al., 2018; Faulds & Raju, 2021; Moran, 2018; Lebovitz et al., 2021). In pursuit of these benefits, organizations across industries and disciplines, including healthcare (Topol, 2019), manufacturing (Lee et al., 2018), law (Surden, 2019), and academia (Hosseini et al., 2024), have adopted AI tools to improve decision-making and operational efficiency.

Despite the promises and claims that AI tools can improve decision-making and efficiency, several companies have struggled to reap the anticipated benefits (Fountaine et al., 2019), primarily due to the complexities of understanding AI's technical intricacies and the challenges of integrating these tools into existing organizational systems and workflows (Chowdhury et al., 2023; Cruz, 2024; Lebovitz et al., 2022). Moreover, AI systems are not neutral computational tools; rather, their effectiveness and viability depend on how they are trained, how they process information, and how they are deployed within existing organizational systems (Cruz, 2024; Lebovitz et al., 2022; Lebovitz et al., 2023; Stahl, 2021). Furthermore, deep learning AI systems are not static; their abilities are continuously shaped by human decisions, biases embedded in training data, and the transparency of their decision-making processes (Koh et al., 2024; Lebovitz et al., 2023). Altogether, these complexities underscore the necessity for organizations and their members to adopt a structured approach to AI evaluation—one that moves beyond simple classification frameworks (e.g., rule-based systems, machine learning, and deep learning; Russell & Norvig, 2021) and instead considers the AI's underlying data sources, decision-making transparency, real-world applicability, and ethical implications.

We propose that a revised approach to evaluating AI tools can offer valuable insights for organizational decision-makers, despite the challenges posed by the ever-evolving nature of these technologies. This is not to suggest that organizations aban-

https://doi.org/10.1515/9783111595306-006

don existing assessment practices, such as predefined performance metrics or regulatory checklists for their current AI-enabled systems. Instead, we contend that the practice of formulating thoughtful and targeted evaluation questions plays a vital role in effective AI assessment and implementation. Accordingly, the approach proposed here focuses on equipping readers with a structured, practical process for assessing AI capabilities, asking the right questions, and making informed decisions about tools' integration and adoption. It is specifically designed for individuals with domain expertise who may not be AI specialists but seek to evaluate AI tools and their integration into organizational processes more effectively. To support this process, each step is supplemented with key questions that help readers critically examine AI's functionality, data sources, transparency, and alignment with organizational needs. To promote accessibility, we intentionally avoid overly technical discussions of AI methodologies that might discourage readers, including organizational decision-makers, from engaging with this framework. Where deeper technical understanding is necessary, we provide references to relevant sources, allowing readers to explore specific AI methodologies as needed without overwhelming the core evaluation process.

6.2 Establishing a framework for the evaluation of AI capabilities

Technology and organizations are fundamentally co-constitutive; they shape one another over time (Orlikowski, 2007, 2010; Leonardi, 2013). In practice, organizations recalibrate their practices to harness the capabilities of newly implemented technologies. Conversely, when confronted with a technology's limitations, organizations may pursue modifications, refining the technology's capabilities to better align with their strategic goals (Leonardi, 2011). These reciprocal adjustments support a co-evolution, during which neither the technology nor those who interact with it necessarily remains the same.

We suggest that organizations that actively engage with this interplay—that is, those that recognize the interdependent nature of human agency and new technologies—are well positioned to maximize the value of their AI investments. Indeed, an organization's competitive advantage often stems not simply from possessing difficult-to-replicate assets but from its ability to continuously adapt, integrate, and evolve technologies in response to environmental changes (Teece, 2007). Organizations that embrace this conceptual framing may therefore respond more dynamically to the emergent risks and opportunities of AI-enabled systems rather than be constricted by their current capabilities.

As software engineers rapidly develop AI-powered applications and organizational leaders integrate them at an unprecedented pace, many workplaces are poised to redefine their policies and practices. However, current implementations risk ossify-

ing workplace norms, making them resistant to revision or, at the very least, leading to increasingly homogenized practices over time (Endacott & Leonardi, 2024). This concern is both operational and strategic; the rapid deployment of AI often outpaces the development of the evidence-based theory that should inform emerging AI policies in organizations (Leonardi, 2023). Thus, we question whether early implementation efforts have already fundamentally shaped, or even constrained, how organizational members negotiate AI's role in their work.

We further posit that this conceptual framing emboldens organizational decision-makers to evaluate AI tools iteratively. Through such ongoing evaluation, decision-makers can more effectively discern not only when but also how existing practices should be restructured to accommodate new tools—and conversely, when and how the technology itself should be refined to optimally support both short- and long-term objectives. These informed decisions may profoundly shape how organizational members reframe their work tasks and evolve their expectations regarding what the technology does (and does not) afford them.

Practically, we suggest that AI-enabled systems—whether already implemented or anticipated—are best evaluated through ongoing, thoughtful inquiry into the forces that shape them. This includes, but is not limited to, evaluations of training data sets, the organization's operational environment, and interactions between users and the tools in question. The ability to frame well-posed questions is not merely an academic exercise; it shapes what is measured, how outcomes are interpreted, and what actions follow (Bouchard, 1976). More critically, the practice of asking such critical questions can determine whether the evaluation of AI capabilities remains a passive, assumption-driven exercise or develops into an active, reflexive process that evolves alongside both the technology and its users.

6.3 Classifying AI based on capabilities

Organizational members seeking to evaluate AI tools for potential applications within their organizations must first understand their capabilities. Gaining insights into an AI tool's primary functions allows decision-makers to determine how AI algorithms can be effectively deployed within specific contexts to achieve desired outcomes (Kaplan & Haenlein, 2020). For instance, AI algorithms used for fraud detection in banking transactions differ from those employed for anomaly detection in industrial equipment and further differ from those designed for cybersecurity threat detection. While all these AI systems identify irregular patterns, their objectives, input data, and decision-making processes are uniquely tailored to their respective domains—financial transactions, machinery performance, and network activity. Additionally, AI capabilities are not uniform; some tools specialize in perception, others in analysis, decision-making, or interaction. By classifying AI tools based on their core capabilities,

organizations can make more informed decisions about their implementation and alignment with operational needs (Schmid et al., 2021).

Inspired by and aligned with human cognitive capabilities, AI technologies can be broadly categorized based on their core functions: *Sensing*, *Processing*, *Acting*, and *Communicating* (Schmid et al., 2021). Following Bloom's (1956) taxonomies—which categorize human abilities into the affective domain (processing sensory impressions), the psychomotor domain (muscle control and coordination), and the cognitive domain (reasoning, problem-solving, and knowledge acquisition)—Schmid and colleagues (2021) suggest that while some AI tools focus on perceiving and gathering data, others analyze information, detect patterns, execute automated actions based on predefined rules, or facilitate interactions through text, voice, or visual outputs. Although this classification is broad, it serves as a heuristic framework to help organizational members identify the functional boundaries of an AI tool and structure their capabilities in a way that fits with their specific use case. The following sections provide a detailed discussion of these four categories, outlining key questions and considerations for their evaluation and implementation.

6.3.1 Sensing AI: Perceiving the Environment

AI systems that perceive and gather data from the environment can be categorized into systems that have *sensing* capability (Schmid et al., 2021). These tools replicate human sensory functions—seeing, hearing, smelling, tasting, and touching—by leveraging AI to interpret and process sensory inputs. For example, AI-powered vision detects objects in autonomous vehicles using LiDAR (Waymo; Nadikattu, 2019), speech recognition transcribes spoken language into text (Google Speech-to-Text; Chiu et al., 2018), and olfactory sensors identify chemical compositions in the air (electronic noses; Kim et al., 2022). Similarly, AI-driven chemical sensing utilizes graphene-based sensors with machine learning to detect food composition, spoilage, and fraud (Pannone et al., 2024), while tactile sensors, like Meta Digit 360, achieve human precision, detecting forces as small as one millinewton and capturing fine spatial details (Lambeta et al., 2024).

A fundamental prerequisite for sensing AI applications is their ability to process and extract meaningful patterns from raw sensor data. While the reliability and accuracy of AI in capturing and interpreting sensory data are crucial, effective implementation extends beyond AI's capabilities—it also requires evaluating whether the organization can provide the necessary sensory input in a structured, reliable, and scalable manner. Researchers emphasize that the success of AI-driven sensing technologies depends heavily on the availability, quality, and consistency of incoming data (Chen et al., 2024). For instance, AI-driven computer vision tools require high-resolution image streams, speech recognition models depend on clean, labeled audio datasets, and olfactory sensors need precise chemical composition data for accurate classification. Organ-

izations considering AI solutions with sensing capabilities must, therefore, assess their ability to supply, process, and maintain the required input data along with AI's functional capacity. Key questions for evaluation include:

a) How reliably and accurately can the AI system capture and interpret sensory input under varying operational conditions, and what are its limitations?
b) What type of sensory input does the AI require, and in what format must the data be provided?
c) Does the organization's existing infrastructure support the collection, storage, and processing of this data while ensuring compliance with privacy and security regulations?
d) Can the AI generalize effectively across different environments, or does it require frequent recalibration and maintenance?
e) What are the system's scalability constraints, including energy consumption, hardware dependencies, and long-term computational demands?

6.3.2 Processing AI: Understanding and Analyzing Data

AI systems that analyze data, detect patterns, and generate insights fall under the *processing* AI category (Schmid et al., 2021). Unlike sensing AI, which primarily collects data from the environment, processing AI interprets, classifies, and extracts meaningful relationships from structured and unstructured data. These systems rely on complex analysis, pattern recognition, and decision-making, typically employing machine learning algorithms, deep neural networks, natural language processing (NLP), and statistical modeling to make sense of vast datasets (Brown et al., 2020; Buczak & Guven, 2016; Jiang et al., 2017). Processing AI applications perform four core functions: detecting and recognizing patterns, deriving meanings and relations, predicting future outcomes, and optimizing decision-making. Schmid and colleagues (2021) further emphasize that distinct methods—e.g., clustering algorithms for pattern recognition, regression models for predictive analytics, and reinforcement learning for decision optimization—are required for these different processing functions.

Pattern recognition AI uncovers recurring structures and anomalies in massive datasets and assists in fraud detection, cybersecurity, and biometric identification. Palantir Gotham, for example, applies machine learning to financial and law enforcement data to expose fraudulent and illegal activities (Munn, 2017). AI systems that extract meaning and relationships process unstructured data (i.e., text, speech, and images) to generate insights. OpenAI's GPT-4 exhibits this capability, using large language models (LLMs) to interpret language, generate human-like responses, and facilitate automated summarization and machine translation (OpenAI, 2023). Predictive modeling AI anticipates future trends by analyzing historical data—a capability demonstrated by Google DeepMind's AlphaFold, which predicts protein folding structures to accelerate drug discovery (Jumper et al., 2021). Finally, decision-optimization AI refines human reasoning

by incorporating logic, context, and adaptive learning. Palantir's AI Platform, for instance, integrates large-language models and real-time data visualizations to support patient and staffing allocation (Tampa General Hospital, 2024).

A key requirement for processing AI is access to high-quality, well-structured data that aligns with its training, processing method, and intended function. However, real-world implementation of these applications often encounters challenges such as data biases, contextual misinterpretations, environmental sensitivity, and computational limitations, which can undermine reliability and effectiveness (Schmid et al., 2021). Additionally, processing AI has function-specific requirements: predictive AI depends on forecasting accuracy and bias mitigation, pattern recognition AI requires monitoring for false positives and ethical concerns, meaning-derivation AI must accurately interpret linguistic and contextual nuances, and decision-optimization AI must function within real-world constraints. Thus, effective implementation requires organizations to assess not just AI's technical capabilities but also its integration with an organization's dynamic operational demands, ethical standards, and data infrastructure. Organizations considering processing AI must evaluate the following:

a) What type of data does the AI system require, and in what format should it be provided?
b) How does the AI analyze, interpret, and process both structured (e.g., databases) and unstructured (e.g., text, images) data?
c) What assumptions does the AI make when identifying patterns? How well does it adapt to new or unseen data, and what mechanisms exist to handle inconsistencies or biases?
d) What are the limitations of the AI's reasoning and pattern-recognition capabilities in complex, ambiguous scenarios?
e) What computational and infrastructure requirements must be met to ensure optimal performance and scalability?

6.3.3 Acting AI: Automating Task Execution

Acting AI tools execute automated tasks, implement decisions, and control physical or digital systems, making them particularly effective for workflow automation and minimizing human intervention in repetitive tasks. These systems respond to processed data by initiating actions, whether through robotic process automation (RPA), rule-based execution, intelligent control systems, or dynamic system adjustments (Hildesheim et al., 2023; Schmid et al., 2021). Depending on how they execute actions and their level of autonomy in operational decision-making, acting AI tools can be classified into four functions: physical actuation, digital execution, interactive response, and autonomous decision-making (Schmid et al., 2021).

Physical actuation AI controls and manipulates physical objects in manufacturing, robotics, and autonomous vehicles. Boston Dynamics' Spot, an AI-powered quadruped

robot, assists in industrial inspections, hazardous environment monitoring, and logistics automation (Chang & Park, 2024). Digital execution AI, such as UiPath's Robotic Process Automation (RPA) platform, streamlines administrative processes by automating data entry, invoice processing, and IT operations, reducing human error and workload (Aguirre & Rodriguez, 2017). Interactive response AI, like Amazon Alexa, uses natural language processing (NLP) and machine learning to personalize interactions commonly used in customer service chatbots, virtual assistants, and recommendation engines (Hoy, 2018). Finally, autonomous decision-making AI operates independently in dynamic environments, making real-time decisions in applications such as self-driving vehicles (e.g., Tesla's Full Self-Driving AI; Bojarski et al., 2016) and algorithmic trading (e.g., DeepTrader AI; Wang et al., 2021).

Given that acting AI systems directly influence physical or digital actions, organizations must carefully evaluate (a) their decision-making mechanisms, (b) the level of human oversight required, and (c) their adaptability in changing conditions. Additionally, different types of acting AI require task-specific considerations. For instance, evaluating physical actuation AI involves assessing precision, adaptability, and operational safety in human environments. In contrast, digital execution AI must be examined for scalability, compliance with data security regulations, and integration with existing systems. Similarly, interactive AI should be assessed for conversational accuracy, personalization, and responsiveness, while autonomous AI must be evaluated in terms of reliability, decision-making transparency, and ability to handle unexpected scenarios. Overall, acting AI tools should be evaluated for their efficacy, alignment with organizational objectives, and potential risks. Key questions to ask during the evaluation process include:

a) What criteria does the AI use to determine its actions, and does it require human oversight or operate fully autonomously?
b) Can the AI adapt to unforeseen conditions, or does it strictly follow predefined rules?
c) What safeguards ensure the AI executes actions safely, accurately, and ethically, particularly in human environments?
d) How does the AI integrate with existing workflows, hardware, or software ecosystems while maintaining compliance with security and regulatory requirements?
e) What mechanisms are in place for monitoring, auditing, and adjusting AI-driven actions to ensure long-term reliability and performance?

6.3.4 Communicating AI: Facilitating Human-AI Interaction

AI systems that generate, interpret, and respond to human input through text, speech, or visual outputs fall under the *communicating* AI category. These tools are integral to customer service, education, accessibility, and content generation, encompassing chatbots, virtual assistants, and AI-driven creative platforms (Schmid et al., 2021). Unlike

acting AI, which focuses on task execution, communicating AI prioritizes information exchange and adaptive engagement, enabling automated customer support, real-time language translation, and personalized content recommendations. Depending on how AI processes and delivers information, communicating AI can be categorized into text-based interaction, speech-based conversation, visual representation, and multimodal communication—each serving distinct interaction needs across various applications (Schmid et al., 2021).

Text-based interaction AI (e.g., OpenAI's ChatGPT) leverages large language models (LLMs) to generate human-like text, answer complex queries, and assist with writing tasks (Brown et al., 2020). Speech-based AI systems, such as Google Assistant, use speech recognition and machine learning to process verbal inputs, respond to commands, and facilitate conversations (Chiu et al., 2018). Visual representation AI tools like DALL·E transform textual or numerical descriptions into highly detailed images and support various marketing, design, and storytelling applications (Ramesh et al., 2022). Lastly, multimodal communication AI integrates multiple input and output modes, including text, speech, and visual elements, to enhance interactivity. Microsoft's Seeing AI, for instance, assists visually impaired users by describing surroundings, reading text aloud, and recognizing faces through a combination of computer vision and speech synthesis (Microsoft, 2023).

Given that communicating AI systems mediate human-AI interactions, their successful deployment hinges on (a) accuracy and reliability, (b) adaptability to diverse user needs, and (c) ethical and transparency considerations. Text- and voice-based AIs must be contextually aware and linguistically adaptive to provide accurate, relevant, and culturally sensitive responses across diverse demographics; otherwise, misinterpreting language clues can lead to miscommunication or reinforce biases, particularly in customer service and automated content generation (Bender et al., 2021). Similarly, visual AI must be historically and culturally accurate to avoid perpetuating myths or inaccuracies. For example, AI-generated images have been criticized for distorting historical events, such as inaccurately depicting Black Confederate soldiers (Levin, 2023). Lastly, a crucial evaluation criterion is how these systems integrate with human workflows. Rather than replacing employees, AI chatbots should complement human agents by handling routine inquiries while escalating complex cases for human oversight. In light of this discussion, prior to deploying communicating AI tools, organizations must consider the following:

a) How does the AI generate responses, and does it provide explanations for its outputs to ensure transparency?
b) How does the AI maintain accuracy, minimize bias, and prevent misinformation or harmful content in its interactions?
c) Can the AI adapt to diverse users, including those with different communication styles, cultural backgrounds, and accessibility needs?

d) What medium of communication (text, voice, visual, or multimodal) best aligns with the organization's needs, and how well does the AI integrate into existing workflows?
e) Does the AI consistently maintain tone, accuracy, and user engagement across various contexts and applications?

By classifying AI tools based on their four core capabilities—sensing, processing, acting, and communicating—organizations can develop a structured understanding of how different AI applications function and align with their operational needs. While advancements in AI may lead to hybrid models that blur these distinctions (Schmid et al., 2021), these core capabilities remain critical for understanding AI's role in various contexts. The classification framework presented here provides a heuristic starting point for assessing AI applications, determining their suitability, and identifying potential limitations. Additionally, the evaluation questions outlined in each category may help organizations assess not only AI's technical capabilities but also the complementary resources, infrastructure, and strategies required for successful adoption.

While this classification framework offers a useful heuristic for categorizing AI tools, it assumes AI capabilities exist independently of their training data and deployment context (Koh et al., 2024; Lebovitz et al., 2023). In reality, AI systems do not simply "sense" or "act" in isolation; their performance depends on the quality of training data, operational environment, and user interactions (Stahl, 2021). For example, two AI models classified as "processing" tools may perform differently based on their training data; one trained on diverse, high-quality datasets may generalize well, while another trained on biased or incomplete data may yield misleading outcomes. AI capabilities are actively shaped by training data rather than reflecting an objective reality (Koh et al., 2024). Therefore, AI classification serves as a starting point but must be supplemented with a deeper analysis of how systems are trained and validated.

6.4 Evaluating AI's training data, transparency, and performance

One of the most determinative factors of an AI tool's performance is the *training dataset*, referred to as ground truth (Lebovitz et al., 2021). AI models learn from historical examples, and if that training data contains biases, errors, or unrepresentative samples, the AI will internalize and replicate those flaws (Lebovitz et al., 2023). This can lead to discriminatory or misleading outcomes, particularly in applications with social or ethical implications (Koh et al., 2024). For instance, Amazon's AI-powered hiring tool was designed to automate candidate selection, but it inadvertently discriminated against female applicants because it was trained on historical hiring data dominated by male applicants. The system downgraded resumes that included the

word “women’s” (e.g., “Women’s Chess Club Captain”) because past hiring decisions favored men in technical roles (Dastin, 2022).

Most AI systems rely on ground truth labels—predefined “correct” answers in training data—to learn patterns and make predictions (Lebovitz et al., 2023). However, these labels are only as accurate as the datasets they originate from. When they reflect historical biases, inconsistencies, or incomplete information, they distort AI performance in real-world applications (Koh et al., 2024). For example, dermatology AI models trained primarily on lighter skin tones have struggled to diagnose skin cancer in patients with darker skin (Rezk et al., 2022). In such cases, the ground truth labels used for training may not fully capture how conditions manifest in different populations. If melanoma images used for training primarily feature lighter skin, the AI may fail to recognize how the disease manifests on darker tones, leading to misdiagnoses. Understanding how training data is selected, labeled, and validated is therefore crucial for assessing an AI system’s accuracy and reliability.

A persistent challenge in AI adoption is the *lack of transparency*; many AI systems function as “black boxes” and make decisions that end-users cannot easily interpret. Without transparency, users may either blindly trust AI outputs or dismiss them due to uncertainty, particularly in high-stakes fields like healthcare and finance (Bauer et al., 2021; Lebovitz et al., 2022). In AI-driven medical diagnostics, for example, radiologists expressed reluctance to use AI recommendations when the system failed to provide clear reasoning for its outputs (Lebovitz et al., 2022). This lack of interpretability extends beyond user confidence: opaque AI models also create accountability gaps where it is unclear who is responsible for an AI-generated error (Bauer et al., 2021). If AI decisions cannot be audited or contested, organizations risk operational failures with no clear path for corrective action.

Explainability is not just about understanding how AI works: it is about ensuring AI-driven decisions align with ethical, regulatory, and practical expectations. In high-risk domains, transparency mechanisms such as explainable AI (XAI) models can help users interpret why a certain decision was made rather than just providing an output (Arrieta et al., 2020; Bauer et al., 2021). However, transparency alone does not guarantee responsible AI usage; organizations must also establish clear governance structures defining who is accountable for AI-driven actions and how faulty outputs can be corrected. Thus, evaluating AI transparency requires assessing whether outputs can be audited, whether users can challenge AI decisions, and whether decision-making processes are interpretable to end-users (Lebovitz et al., 2023).

Organizations often rely on technical *performance metrics* such as accuracy, precision, and recall to evaluate AI tools. However, these metrics do not always reflect real-world performance (Lebovitz et al., 2023). AI systems may achieve high accuracy in controlled test settings yet fail in practice due to unforeseen variables. For example, an AI model trained for fraud detection may flag transactions effectively in a test but misidentify legitimate transactions in live banking operations due to subtle contextual differences. In addition, accuracy does not equate to reliability—some AI sys-

tems generate highly convincing but incorrect outputs, a phenomenon known as "artificial hallucination" (Alkaissi & McFarlane, 2023). For example, in an analysis of ChatGPT's scientific writing, researchers found that the model fabricated citations and plausible-sounding but incorrect explanations (Alkaissi & McFarlane, 2023).

Beyond accuracy, evaluating AI performance requires assessing its adaptability, robustness, and reliability in real-world conditions. AI models trained in controlled environments often struggle when exposed to operational complexities they were not designed to handle. For example, an AI-driven sepsis prediction tool performed well in clinical trials but failed in hospitals with different patient demographics and data collection methods (Sendak et al., 2020). Similarly, autonomous driving AI may perform flawlessly in structured test simulations but misinterpret unexpected road conditions and increase accident risks. To ensure AI tools function reliably beyond lab settings, organizations must test their ability to generalize, identify failure points, and provide explanations for their outputs. Without this evaluation, AI systems may appear successful in development but introduce unforeseen risks in deployment.

A thorough evaluation of AI systems should encompass an analysis of the training data's integrity, the transparency of the model's decision-making processes, and a holistic assessment of performance that goes beyond conventional metrics. Biases in training data must be actively identified and addressed, explainability must be prioritized, and AI performance should be evaluated in practical settings. To support this evaluation process, Table 6.1 presents key questions organizational members must consider when evaluating AI systems.

Table 6.1: Key Questions for Evaluating AI Training Data, Transparency, and Performance.

Evaluation Factor	Key Questions
Training Data	– What is the source of the training data? – Does the data contain historical biases? – Were diverse perspectives included? – How was ground truth defined?
Transparency	– Can users interpret the AI's reasoning? – Is there a mechanism to audit or challenge AI outputs? – Who is accountable if the AI makes an error?
Performance	– Does the AI perform well outside controlled conditions? – Is the AI prone to bias or hallucinations? – Can the AI's performance be interpreted and understood?

6.5 Evaluating AI capabilities in context

Inspired by the recommendations portrayed here to review an AI's capabilities iteratively, consider the following scenario:

> At the behest of its administrators, a leading research university implemented an AI tool to assist a faculty-led committee in its allocation of research funding among its members. The tool, designed to interpret grant proposals, citation records, and previous funding outcomes, recommended to committee members whether a proposed project should receive financial support from the institution. University administrators foresaw the tool rendering efficient, bias-free, and prudent funding decisions. But shortly after being introduced to the tool, faculty raised concerns about the AI's influence on the university's research priorities.
>
> As anticipated, the faculty committee initially appreciated the AI's ability to process grant proposals and their myriad supplemental data. Yet when early-career faculty voiced concerns that their proposals were routinely rejected, the committee realized a pattern in the tool's recommendations: research proposals with methodologies and topics that aligned with existing institutional priorities received higher rankings, and novel, interdisciplinary, or emerging research areas received lower rankings. Indeed, early-career faculty were those proposing the unconventional approaches, which the tool deprioritized. Upon making this discovery, the committee also reflected on a shift in how it had managed funding discussions: the AI's rankings had subtly narrowed the committee's definition of fundable research. In fact, the committee realized that it had grown hesitant to challenge the tool's recommendations, fearing that subjective judgments might appear less rigorous than supposedly data-driven rankings.
>
> The committee thus questioned whether the tool had been reinforcing past funding patterns and necessarily overlooking those breaking new ground in their fields. The committee compared AI-generated funding recommendations with historical funding decisions made through traditional peer review. When the results of this inquiry affirmed the committee's suspicions, the committee implemented several key changes. First, it recommended refinements to the AI's training data, incorporating qualitative assessments from recent faculty review panels. This ensured that funding recommendations no longer relied solely on metrics like citation counts but considered other criteria, such as originality. Then, the committee moved to allow applicants to contest AI-generated rankings and to advocate for their work. Finally, the committee mandated that the outcomes of this and any forthcoming AI tool serve as a supplemental input among many in awarding funding.

This scenario underscores the interplay between technology and its users, illustrating that an AI-enabled system is neither neutral nor static. In this case, the tool's recommendations are revealed to reflect the biases and assumptions embedded in its training data. The tool's influence is also shaped by the committee's interaction and interpretation of its results. But the committee is quick to change course; faculty actively define the problem at hand, interrogate the tool's outputs, and steer its use toward supporting evolving institutional goals rather than allowing it to passively steer their institution's research landscape.

The committee's eventual course corrections indeed mirror the very practices encouraged by this chapter's framework. The framework put forth may well have served the committee in addressing their concerns in advance. First, by considering what type of AI was being deployed—in this case, a Processing AI tasked with evaluat-

ing unstructured writing alongside metrics—the committee might have clarified whether such a tool was suitable for interpreting interdisciplinary research proposals. Second, by demanding greater transparency from the tool by asking, for example, how the AI was trained, the committee may have spotted its narrowing effect on what constitutes fundable research far sooner. Finally, had the committee devised a clear method for evaluating the AI's performance in context, misalignments between AI-generated recommendations and the university's values would be evident early in the tool's implementation.

Just as effective risk management depends on routine exercises, scenario testing, and iterative adjustments to maintain resilience (Engemann & Engemann, 2022), the responsible integration of this AI tool requires regular audits as well as the engagement of its users. Assuming the committee sustains its efforts beyond the given scenario, the AI is poised to become more responsive to users' judgment. By employing the proposed framework that emphasizes critical inquiry of the tool by its users, the university ensures that the AI remains a constructive force in its research funding decisions. Moreover, the university's broader commitments, such as fostering scholarly innovation, are likely to emerge more clearly as institutional priorities.

6.6 Conclusion

We advance a structured yet adaptable framework for evaluating AI capabilities, emphasizing that successful implementation requires thoughtful, iterative inquiry. By classifying AI-enabled systems and tools according to their core functions—sensing, processing, acting, and communicating—and embedding this classification within broader considerations of transparency, organizational members are optimally prepared to identify the assumptions, risks, and limitations that shape AI performance. Moreover, this framework positions users as active participants in interpreting and refining how these tools are integrated into their work. The promise of AI is not realized through adoption alone; rather, through continuous reflection on what the technology affords, what it obscures, and how it reshapes decision-making, organizations and their members are better equipped to harness its potential.

References

Aguirre, S., & Rodriguez, A. (2017). Automation of a business process using robotic process automation (RPA): A case study. In Figueroa-García, J., López-Santana, E., Villa-Ramírez, J., & Ferro-Escobar, R. (Eds.). *Applied Computer Sciences in Engineering*, *WEA 2017* (pp. 65–71). Springer. https://doi.org/10.1007/978-3-319-66963-2_7

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, *15*(2), e35179. https://doi.org/10.7759/cureus.35179

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., . . . & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bauer, K., Hinz, O., van der Aalst, W., & Weinhardt, C. (2021). Expl(AI)n it to me–explainable AI and information systems research. *Business & Information Systems Engineering*, *63*, 79–82. https://doi.org/10.1007/s12599-021-00683-2

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. https://doi.org/10.1145/3442188.3445922

Bloom, B. (1956). *Taxonomy of educational objectives, Vol. 1*: *Cognitive domain*. David Mckay.

Bojarski, M., Testa, D. D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., . . . & Zhang, X. (2016). End-to-end learning for self-driving cars. *arXiv Preprint*, arXiv:1604.07316. https://arxiv.org/abs/1604.07316

Bouchard, T. J. (1976). Field research methods: Interviewing, questionnaires, participant observation, systematic observation, unobtrusive measures. In M. D. Dunnette (Ed.), *Handbook of Industrial Organizational Psychology* (pp. 363–413). Rand McNally.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cybersecurity intrusion detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Bughin, J., Hazan, E., Lund, S., Dahlström, P., Wiesinger, A., & Subramaniam, A. (2018). Skill shift: Automation and the future of the workforce. *McKinsey Global Institute*. https://www.mckinsey.com/featured-insights/future-of-work/skill-shift-automation-and-the-future-of-the-workforce

Chang, J., & Park, J. (2024). Design of a robot system for surveillance and anomaly detection in industrial environments. In *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)* (pp. 2229–2234). IEEE. https://doi.org/10.1109/ICTC62082.2024.10827327

Chen, L., Xia, C., Zhao, Z., Fu, H., & Chen, Y. (2024). AI-Driven Sensing Technology: Review. *Sensors*, *24*(10), 2958. https://doi.org/10.3390/s24102958

Chiu, C. C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., . . . & Bacchiani, M. (2018). State-of-the-art speech recognition with sequence-to-sequence models. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4774–4778. https://doi.org/10.1109/ICASSP.2018.8462105

Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., & Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through AI capability framework. *Human Resource Management Review*, *33*(1), 100899. https://doi.org/10.1016/j.hrmr.2022.100899

Cruz, I. F. (2024). Expert-AI pairings: Expert interventions in AI-powered decisions. *Information and Organization*, *34*(4), 100527. https://doi.org/10.1016/j.infoandorg.2024.100527

Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics* (pp. 296–299). Auerbach Publications.

Endacott, C. G., & Leonardi, P. M. (2024). Artificial intelligence as a mechanism of algorithmic isomorphism. In *Research Handbook on Artificial Intelligence and Decision Making in Organizations* (pp. 342–358). Edward Elgar Publishing. https://doi.org/10.4337/9781803926216.00029

Engemann, K. N., & Engemann, K. J. (2022). Integrating dynamic modeling solutions towards a resilience model. In *Organizational Risk Management: Managing for Uncertainty and Ambiguity* (pp. 97–110). DeGruyter.

Faulds, D. J., & Raju, P. (2021). The work-from-home trend: An interview with Brian Kropp. *Business Horizons*, *64*(1), 29–35. https://doi.org/10.1016/j.bushor.2020.10.005

Fountaine, T., McCarthy, B., & Saleh, T. (2019). Building the AI-powered organization. *Harvard Business Review*, *97*(4), 62–73. https://hbr.org/2019/07/building-the-ai-powered-organization

Hildesheim, W., Holoyad, T., & Schmid, T. (2023). Machine learning in AI factories – five theses for developing, managing, and maintaining data-driven artificial intelligence at large scale. *it – Information Technology*, *65*(4–5), 218–227. https://doi.org/10.1515/itit-2023-0028

Hosseini, M., Rasmussen, L. M., & Resnik, D. B. (2024). Using AI to write scholarly publications. *Accountability in Research*, *31*(7), 715–723. https://doi.org/10.1080/08989621.2023.2168535

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, *37*(1), 81–88. https://doi.org/10.1080/02763869.2018.1404391

Jiang, F., Jiang, Y., Zhi, H., . . . & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., . . . & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kaplan, A., & Haenlein, M. (2020). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons*, *63*(1), 37–50. https://doi.org/10.1016/j.bushor.2019.09.003

Kim, C., Lee, K. K., Kang, M. S., Shin, D. M., Oh, J. W., Lee, C. S., & Han, D. W. (2022). Artificial olfactory sensor technology that mimics the olfactory mechanism: A comprehensive review. *Biomaterials Research*, *26*(1), 40. https://doi.org/10.1186/s40824-022-00287-1

Koh, J. (2024). Beyond data collection: Examining artificial intelligence data creation in organizations. *Academy of Management Proceedings*, *2024*(1), 17867. https://doi.org/10.5465/AMPROC.2024.92bp

Lambeta, M., Wu, T., Sengul, A., Most, V. R., Black, N., Sawyer, K., . . . & Calandra, R. (2024). Digitizing touch with an artificial multimodal fingertip. *arXiv Preprint*, arXiv:2411.02479. https://arxiv.org/abs/2411.02479

Lee, J., Davari, H., Singh, J., & Pandhare, V. (2018). Industrial Artificial Intelligence for industry 4.0-based manufacturing systems. *Manufacturing Letters*, *18*, 20–23. https://doi.org/10.1016/j.mfglet.2018.09.002

Levin, K. M. (2023, February 9). The Black Confederate myth: Artificial intelligence and the perpetuation of a myth. *Civil War Memory*. https://kevinmlevin.substack.com/p/the-black-confederate-myth-artificial

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI ground truth really "true"? The dangers of training and evaluating AI tools based on experts' know-what. *Management Information Systems Quarterly*, *45* (3b), 1501–1525. https://doi.org/10.25300/MISQ/2021/16564

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization science*, *33*(1), 126–148. https://doi.org/10.1287/orsc.2021.1549

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2023). The No. 1 question to ask when evaluating AI tools. *MIT Sloan Management Review*, *64*(3), 27–30. https://sloanreview.mit.edu/article/the-no-1-question-to-ask-when-evaluating-ai-tools/

Leonardi, P. M. (2013). Theoretical foundations for the study of sociomateriality. *Information and Organization*, *23*(2), 59–76. https://doi.org/10.1016/j.infoandorg.2013.02.002

Leonardi, P. M. (2011). When flexible routines meet flexible technologies: Affordance, constraint, and the imbrication of human and material agencies. *MIS Quarterly*, *35*(1), 147–167. https://doi.org/10.2307/23043493

Leonardi, P. M. (2023). Affordances and agency: Toward the clarification and integration of fractured concepts. *MIS Quarterly*, *47*(4), ix-xx.

Microsoft. (2023). *Seeing AI: Talking camera app for the blind community*. https://www.microsoft.com/en-us/seeing-ai

Moran, G. (2018). This artificial intelligence won't take your job, it will help you do it better. *Fast Company*. https://www.fastcompany.com/90253977/this-artificial-intelligencewont-take-your-job-it-will-help-you-do-it-better.

Munn, L. (2017). Seeing with software: Palantir and the regulation of life. *2*(1). https://doi.org/10.17613/5cq8-7w72

Nadikattu, R. R. (2019). New ways in artificial intelligence. *International Journal of Computer Trends and Technology*, *67(12)*, 89–94. https://dx.doi.org/10.2139/ssrn.3629063

OpenAI. (2023). *GPT-4 technical report*. https://openai.com/research/gpt-4

Orlikowski, W. J. (2007). Sociomaterial practices: Exploring technology at work. *Organization Studies*, *28 (9)*,1435–1448. https://doi.org/10.1177/0170840607081138

Orlikowski, W. J. (2010). The sociomateriality of organizational life: Considering technology in management research. *Cambridge Journal of Economics*, *34(1)*, 125–141. https://doi.org/10.1093/cje/bep058

Pannone, A., Raj, A., Ravichandran, H., . . . & Das, S. (2024). Robust chemical analysis with graphene chemosensors and machine learning. *Nature*, *634*, 572–578. https://doi.org/10.1038/s41586-024-08003-w

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv Preprint*, arXiv:2204.06125. https://doi.org/10.48550/arXiv.2204.06125

Rezk, E., Eltorki, M., & El-Dakhakhni, W. (2022). Improving skin color diversity in cancer detection: Deep learning approach. *JMIR Dermatology*, *5*(3), e39143. https://doi.org/10.2196/39143

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Schmid, T., Hildesheim, W., Holoyad, T., & Schumacher, K. (2021). The AI methods, capabilities and criticality Grid. *Künstl Intell*, *35*, 425–440. https://doi.org/10.1007/s13218-021-00736-4

Sendak, M. P., D'Arcy, J., Kashyap, S., Gao, M., Nichols, M., Corey, K., Berner, E. S., & Balu, S. (2020). A path for translation of machine learning products into healthcare delivery. *npj Digital Medicine*, *3*(1), 1–5. https://doi.org/10.33590/emjinnov/19-00172

Stahl, B. C. (2021). Ethical issues of AI. In *Artificial intelligence for a better future* (pp. 35–54). *Springer Briefs in Research and Innovation Governance*. Springer. https://doi.org/10.1007/978-3-030-69978-9_4

Surden, H. (2019). Artificial intelligence and law: An overview. *Georgia State University Law Review*, *35*(4), 1305–1337. https://ssrn.com/abstract=3411869

Tampa General Hospital. (2024, June 5). *Tampa General Hospital selects Palantir's AI software to power connected care coordination, extending long-term partnership*. https://www.tgh.org/news/tgh-press-releases/2024/june/tgh-selects-palantir-ai-software-connected-care-coordination

Teece, D. J. (2007). Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance. *Strategic Management Journal*, *28*, 1319–1350. https://doi.org/10.1002/smj.640

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

Wang, Z., Huang, B., Tu, S., Zhang, K., & Xu, L. (2021). DeepTrader: A deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(1), pp. 643–650. https://doi.org/10.1609/aaai.v35i1.16144

Part II: **Applications**

Jeanne Sheehan

7 Deepfake democracy: The rise of AI in the electoral process

7.1 Introduction

Imagazsine picking up the phone just two days before a crucial election, expecting a routine campaign pitch. Instead, the unmistakable cadence of President Joe Biden resounds through the receiver, urging you not to cast your ballot. This was precisely the situation New Hampshire voter Gail Huntley found herself in ahead of the 2024 primary (Khan 2024).[1] The caller ID indicated Kathy Sullivan, a former chair of the state Democratic Party, but the signature phrase— "what a bunch of malarkey"—revealed that Sullivan was not the real speaker. Huntley recognized the voice as Biden's, the candidate she had voted for in the previous election and intended to support again.

Yet something felt off. The voice was genuinely Biden's in tone and timbre, but it urged Democrats to "save your vote for November" rather than participate in the impending primary (Seitz-Wald 2024). Huntley knew there was no plausible reason for an incumbent president to encourage party members to skip an election. Other New Hampshire residents reported receiving similar calls. Some believed the words were authentic but taken out of context; however, investigators later confirmed the audio was not simply edited Biden recordings. Instead, it had been produced by an advanced artificial intelligence (AI) voice-cloning tool that replicated the president's vocal patterns with striking accuracy (Bond 2024; Khan 2024; Swenson 2024).

Further inquiry revealed that political consultant Steven Kramer orchestrated these calls. The Federal Communications Commission (FCC) charged him with unlawfully making "robocalls … using deepfake, AI-generated voice cloning technology and caller ID spoofing" to mislead voters before the primary (FCC 2024). Kramer faced a proposed $6 million fine for exploiting generative AI to spread misinformation (FCC 2024; Swenson 2024). Although it was not the first time AI-based voice manipulation had been used in politics, the timing—so close to a highly anticipated primary—amplified public alarm about the risks of AI in elections.

Such incidents have fueled growing apprehension among officials, researchers, and the public about the potential for generative AI (GenAI) to disrupt political pro-

1 A change in the Democratic National Committees [DNC's] rules that year made South Carolina – not New Hampshire – the first official primary of the DNC's nominating process. As a result, incumbent President Biden was not one of the twenty-one candidates whose names appeared on the ballot (DeWitt 2024). As a result, Huntley – like almost 80,000 other Democrats from the Granite state, chose to write his name in.

https://doi.org/10.1515/9783111595306-007

cesses.[2] As Professor Hany Faird observed, the main worry is a perfectly timed deepfake that drops "the day before, the day of election, [or] the day after election," possibly igniting chaos with an incendiary video or audio clip (quoted in Bond 2024). Headlines underscored this unease, with Reuters highlighting how "America's 2024 election collides with AI boom," while other outlets warned that AI-generated "deepfakes" could wreak havoc on the presidential race (Ulmer and Tong 2023; Barrabi 2023).

Public opinion data confirmed that such fear was widespread. According to Harry Yaojun Yan et al. (2025), four out of five Americans voiced worries about AI-fueled misinformation prior to the election,[3] while a Pew Research Center survey found that almost four out of ten respondents anticipated AI being used "mostly for bad purposes" during the campaign (Gracia 2024).[4] Another poll by Elon University showed that nearly three-quarters (73%) of Americans considered it "very or somewhat likely" AI would be leveraged to sway electoral results (Elon University 2023).

Against this backdrop, commentators dubbed 2024 "the year of the AI election" (Hsu 2024).[5] There were two primary reasons: first, the availability of AI-enhanced tools—from chatbots to image generators—that facilitated production and dissemination of campaign content, both legitimate and fraudulent. Scholars observed that these resources were employed in contests worldwide, from Bangladesh to Argentina (Saab 2024). Second, 2024 featured a heavy international election calendar, described by International IDEA as a "global elections super-cycle," with more than 70 nations and the European Parliament holding major votes in a single year. Over 1.6 billion voters participated (The 2024 Global 2024), making 2024 the single busiest year for democratic exercises in recorded history.

This convergence of a global electoral season and transformative AI technology raised numerous research questions: What are the most common ways AI has been utilized in campaigning? What are the chief security threats to electoral systems, both domestically and internationally? And what do these developments imply for the future of representative democracy? The study presented in this chapter addresses these issues by investigating (1) the extent and nature of AI-enabled (GenAI) applications, (2) the range of risks (both detrimental and beneficial) GenAI poses in a repre-

2 A Google News search by the author of "New Hampshire primary Biden robocall" (1/1/24-12/15/24) found 1120 articles.

3 Yaojun Yan (2025) also found that these concerns were not correlated with "direct interactions with AI tools. . . instead, news consumption, particularly through television, appeared more closely linked to heightened concerns."

4 Researchers in Europe identified similar concerns, see for instance a study by Carlos Luca de Tena (2024) which found that almost seven out of ten Europeans feared AI manipulation in their elections.

5 Another case of AI use early in the election cycle which received a good deal of media attention was in the summer of 2023 when Florida Governor and candidate for the Republican nomination for president, Ron DeSantis' campaign shared a deepfake of Donald Trump embracing Anthony Fauci (Rehagen 2024).

sentative democracy, and (3) effective practices for mitigating hazards in an increasingly AI-driven political landscape.

7.2 Background

The year 2024 featured an array of scheduled elections around the world, most notably for the European Union (EU) Parliament and the U.S. presidency. Yet in France and the United Kingdom (UK), unexpected developments disrupted the usual electoral timetable. On June 9, 2024, French President Emmanuel Macron dissolved the National Assembly, compelling a snap legislative contest just three weeks later (Le Monde 2024).[6] This followed an equally surprising decision a few weeks prior, when UK Prime Minister Rishi Sunak, speaking under a downpour outside 10 Downing Street, announced an early election in Britain (Paxton et al. 2024).[7] Both announcements arrived barely a month after the EU parliamentary polls, intensifying speculation over how newly accessible AI tools could influence these hastily arranged campaigns.

In response to widespread calls for research into AI's potential role in these elections, a leading team at The Centre for Emerging Technology and Security (CETaS), housed at The Alan Turing Institute, initiated a project entitled "AI Enabled Influence Operations and Election Security." Under the direction of Sam Stockwell, the project produced three comprehensive Briefing Papers between May and November 2024, analyzing developments in the UK, France, EU Parliament, and U.S. elections (Stockwell 2024a, 2024, 2024b).[8] A central concern was that insufficient data on AI-driven threats might amplify public anxieties, eroding confidence in democratic processes (Stockwell et al. 2024b). Although early conclusions suggested no direct evidence that AI-driven misinformation had decisively changed an electoral outcome (i.e., no first-order risk), the findings did indicate mounting second-order risks: public uncertainty regarding what is real or fabricated, heightened prospects for harassment of political figures, and strategic exploitation of AI content by candidates (Stockwell et al. 2024a, 3).

These results parallel those reported by other investigations, including work by Bond (2024), Brack (2024), Denkovski et al. (2024), Schneier and Sanders (2024), and Meta Says AI (2024). Taken together, they reveal that generative AI (GenAI) has yet to overturn election results, yet it can still produce marked effects—both beneficial and detrimental—on public discourse. As is common with second- and third-order threats, the impact

6 Two rounds were held in the French election on June 30 and July 7 respectively.

7 In the UK, Sunak announced the dissolution of parliament in on May 24 and the election took place on July 4.

8 In addition to the work by Stockwell et al (2024a, 2024, 2024b) other work in this area includes research by Kathleen M. Carley, Hoda Heidari, Yonatan Bisk, Sebastien Brack, Ognjan Denkovski, Daniela Alvarado Rincón, Bruce Schneier, Nathan Sanders, and many others (see for instance, Bond 2024, Brack 2024, Denkovski 2024).

often manifests less conspicuously than direct vote manipulation, making it more elusive to detect, mitigate, or counteract. Such long-term erosion of trust may ultimately pose greater danger to democratic integrity than any isolated incident of data tampering.

7.3 Methodology

7.3.1 Research questions

This study explores whether AI-generated content (GenAI) endangers the health of a representative democracy, especially within the context of electoral processes. The following research questions (RQs) guide the investigation:

RQ1: Was GenAI used during the 2024 U.S. election? [Use]
If so:
RQ2: Who created and employed it? [Source]
RQ3: When was it used? [Timing]
RQ4: What are examples of malicious/detrimental and beneficial uses? Which are used most often? [Nature and Type of Use]
RQ5: What can be done to mitigate the risk of detrimental use and encourage beneficial use? [Recommendations]

7.3.2 Conceptualization

Before collecting data, several core terms were defined for clarity (italicized in the RQs above): democracy, representative democracy, AI, GenAI, AI risk, detrimental or malicious use, and beneficial use.

I. Democracy and Representative Democracy
The word “democracy” originates from the Greek demokratia, combining demos (the people) with kratia (rule). In its simplest sense, democracy is “a system of government in which the people rule” (Sheehan 2021, 24–25). A representative or republican democracy involves governance in which the population exerts influence indirectly through elected officials, unlike direct or pure democracies in which the population rules “directly.” Most modern democracies incorporate representative structures, with occasional elements of direct democracy remaining.

II. AI, GenAI, and Types
Numerous definitions of AI exist. For the purposes of this study, AI is treated as “computer systems or machines that use algorithms, data, and computational power to

simulate intelligent behavior and which can perform tasks typically requiring human intelligence. These tasks include problem-solving, learning, reasoning, understanding natural language, recognizing patterns, and decision-making." An additional, equally valid definition offered by Oxford (Artificial Intelligence 2024) portrays AI as "a computer system that can perform complex tasks that would otherwise require human minds—such as visual perception, speech recognition, decision-making, and translation between languages."

Researchers often classify AI into three distinct categories, summarized in Table 7.1 While these definitions vary among sources (Escott 2017; Johnson 2024), they suffice here to distinguish each type's scope and capabilities. This study focuses primarily on GenAI (AGI/strong AI), given its capacity to imitate human-like reasoning and creativity.

Table 7.1: Key types of AI.

Type	Abbreviation & Alternate Terms	Definition	Examples
Narrow/Weak AI	ANI (Narrow or Limited AI)	AI with a specialized focus on tasks, such as language translation or facial recognition	IBM's Watson, traffic management systems, streaming service recommendation engines
General/Strong AI	AGI (Gen AI)	AI capable of performing the full range of intellectual tasks humans can accomplish	Customer service chatbots, deep learning models, self-driving vehicles
Superintelligent AI	ASI	AI surpassing human intellect across every domain	Hypothetical future systems exceeding human problem-solving in all respects

III. AI Risk

AI risk encompasses the likelihood that AI (whether ANI, AGI, or ASI) significantly affects specific contexts—positively or negatively. This research concentrates on the risk GenAI may pose to representative democracies via electoral manipulation. To analyze this, the study distinguishes between detrimental (malicious) and beneficial applications.

IV. Detrimental (Misuse) vs. Beneficial Use

- *Detrimental (Malicious) Use:* Any employment of GenAI intended to undermine democratic procedures, such as spreading disinformation, manipulating voter behavior, or damaging the integrity of elections.
- *Beneficial Use:* Any GenAI application that strengthens democratic processes, such as boosting voter engagement, improving public knowledge, or aiding under-resourced campaigns and communities.

Whether a particular instance of GenAI is detrimental or beneficial depends on its overall impact on the health of the representative system—not merely on whether it favors the user's perspective.

7.4 Data collection and limitations

Data on GenAI usage were gathered from March 1, 2023, to November 4, 2024. This broader timeframe contrasts with narrower windows used by previous research teams, notably Stockwell et al. (2024b),[9] who focused on a six-month period for European and British elections. U.S. campaigns often stretch over considerably longer durations; the 2024 French election concluded in just 30 days (June 9–July 7), and the 2024 UK election spanned six weeks (May 24–July 4). In contrast, a U.S. presidential campaign can last well over a year: in 2024, the victorious candidate, Donald Trump, announced his run on November 15, 2022, nearly two years before Election Day (November 5, 2024). Historical precedents illustrate even longer timelines, such as the 2016 race lasting nearly 600 days, while the 2020 cycle extended beyond three years from the point of the earliest major announcement (Zaino 2024).

Cases of AI-generated content were culled from publicly accessible sources, including news and social media outlets, academic journals, and both online and print publications. Research assistants utilized search engines and AI-based methods to locate and catalog instances of GenAI usage in campaigns. Yet, as with Stockwell et al. (2024, 7), the approach prioritizes identified or "viral" cases—those substantial enough to have drawn notice from journalists or informed observers. Consequently, the study cannot capture cases of AI-driven disinformation or campaign manipulation that remained beneath the visibility threshold required for public or media reporting.[10]

9 In the case of the latter, data was collected for four months, May-August 2024 (Stockwell et al 2024a; Stockwell 2024). It is important to note that their study also looked at use following Election Day, hence the extended time frame. By contrast, the data used in this study was collected through Election Day (November 5, 2024).

10 There is no claim that this is an exhaustive list of cases for this reason and the fact that even cases which did reach some level of virality may have been missed in this examination.

7.5 Findings: Extent and nature of genAI use (RQ1–RQ3)

7.5.1 RQ1 – use

A total of fifty-nine (59) documented cases of GenAI use emerged between February 1, 2023, and November 5, 2024. The earliest instance occurred in mid-February 2023, when a deepfake video circulated on social media depicting President Biden making transphobic remarks (Fact Check 2023). Conversely, the second case presented a starkly different application: the Democratic National Committee (DNC) tested AI to generate initial drafts of fundraising emails, reportedly outperforming purely human-written appeals (Goldmacher 2023).

Nearly two years later—on Election Day—the final identified case centered on a substantial AI-driven bot network on X (formerly Twitter). Researcher Elise Thomas (2024) uncovered this network, which supported Donald Trump's campaign. The network showed varying degrees of sophistication; some accounts blatantly announced their AI origins (e.g., using phrases such as "I am a language model AI created by Open AI") before ultimately being suspended. These bookend examples highlight contrasting ways GenAI was deployed (and, in at least one instance, traced to a known origin) across the long campaign season.[11]

A brief selection of other documented AI uses during the 2024 campaign is shown in Table 7.2 These cases fall between March 2023 and September 2024 and illustrate both international and domestic sources as well as diverse formats—video, photographs, or avatars.

7.5.2 RQ2 – source

Researchers tracked various characteristics for each of the fifty-nine cases, including the origin—or source—of the GenAI deployment. As Table 7.2 hints, some examples originated domestically, while others came from foreign actors.

Table 7.3 groups all fifty-nine incidents into foreign or domestic categories, plus a small group of unknown origin. Nearly three-quarters (73%) were traceable to sources within the United States; one-fifth (20%) were attributed to non-U.S. entities, leaving four cases unassigned.

Determining the origin of these AI-generated efforts is vital for mitigating risk. Some GenAI campaigns might have overlapping objectives, yet whether the source is

11 As Steffen (2024) notes, as of Election Day it was unclear who was behind the bot-network. Moreover, she quotes Thomas in saying that the case underscores one of the "game changing" things about AI which is that "networks like these could be pretty much entirely automated."

Table 7.2: Sample cases of AI use in the 2024 u.s. campaign.

Title	Description	Type	Date	Source	Citation
 Harris left a woman paralyzed in a hit-and-run accident	AI-generated video falsely alleging that Vice President Harris caused a hit-and-run accident, leaving a victim paralyzed 13 years earlier. The footage appeared on a website claiming—falsely—to be a San Francisco news outlet.	Video	September 2024	International (Russian)	Klepper 2024
 Photo of Trump being arrested in NYC	Fabricated image of Donald Trump's arrest purportedly in NYC, produced using Midjourney V5 (an AI text-to-image platform). Shared repeatedly on social media as "parody" by its creator.	Photo	March 2023	International (Eliot Higgins, founder of Bellingcat)	Kochi 2023
 Photos of Trump hugging Dr. Fauci	Video published by Ron DeSantis' campaign containing AI-generated images of Trump embracing Dr. Anthony Fauci.	Photo (in Video)	June 5, 2023	Domestic (DeSantis campaign)	Nehamas 2023
 Little-known candidate beat Biden using AI	Jason Palmer, an obscure Democratic candidate, prevailed over President Biden in the American Samoa caucus with support from a chatbot avatar ("PalmerAI") that ran campaign messaging and outreach.	Avatar	March 2024	Domestic (Palmer's campaign)	Bousquette 2024

Table 7.3: Cases of genai use by source: foreign vs. domestic.

Source	Number of Cases	Percentage
Domestic	43	73%
Foreign	12	20%
Unknown	4	7%
Total (N)	**59**	**100%**

local or international can shape the strategies required for detection and response. In their examination of election-related disinformation, Janjeva et al. (2023) stress that distinct actors (e.g., state governments vs. domestic political campaigns) exhibit different motivations. International state-sponsored disinformation might aim to destabilize foreign politics or influence strategic communication goals, whereas domestic political actors tend to focus on discrediting opponents or personalizing campaign messages.

Table 7.4 further classifies the origin of cases by distinguishing between state, non-state, and campaign-affiliated actors. While only a fifth (20%) of all cases were foreign, nearly all such cases (92%) were linked to state-level sponsorship. Domestic incidents were most frequently tied to political organizations, campaigns, or candidates (93%). Though fewer in number, foreign state-sponsored actions often require broader, resource-intensive countermeasures, such as international collaboration or cooperation with major social media platforms (Janjeva et al. 2023, 21).

Table 7.4: Cases of genai use by source and actor.

Source	Actor	Cases	Percentage of Source	Cumulative % of Total
Foreign	State Actor	11	92%	20%
	Non-State Actor (Individual/Group)	1	8%	2%
Subtotal		12	93%	73%
Domestic	Campaign/Political Org./Candidate	3	7%	5%
	Non-Political Organizations/Individuals	3	7%	5%
Subtotal		43	100%	
Total		59	–	100%

7.5.3 RQ3 – timing

In addition to source, timing plays a key role in any risk management approach. Stockwell et al. (2024a, 10–11) propose separating an election cycle into three phases: (1) Pre-Election (roughly twelve months before Election Day), (2) Polling Period (the 72 hours prior to and including Election Day), and (3) Post-Election (the following one to twelve months). This segmentation allows for pinpointing when different threats tend to arise—essential knowledge for targeting limited resources effectively.

For the U.S. election, cases of GenAI usage in this study were grouped into two primary periods: Pre-Election (PE) and Polling (PO). The PE period (March 1, 2023–August 31, 2024) was further subdivided:

I. Invisible Primary (March 1–December 31, 2023)
II. Primary Season (January 1–June 30, 2024)
III. Conventions (July 1–August 31, 2024)[12]

Likewise, the PO period (September 1–November 5, 2024) was divided between:

I. Early Voting (September 1–November 4, 2024)
II. Election Day (November 5, 2024)[13]

As Table 7.5 shows, 71% of all documented GenAI usage took place during the PE stage. Notably, 40% of PE incidents (17 out of 42) clustered in July and August—the period of major party conventions, when nominees become official.[14] Although only 29% of the total incidents occurred in the PO window, that interval spanned little more than two months. Only a single case was recorded on Election Day itself—a large-scale AI bot network discovered on X, supporting Donald Trump's reelection effort. Just prior to Election Day, three cases surfaced in late October and early November, including an AI-crafted image of a younger Kamala Harris working at a fast-food chain, plus two Russia-linked fake videos targeting immigration and Vice President Harris's spouse (Fact Check 2024; Johnson 2024).

These timing patterns underscore how efforts to deceive or influence voters spiked during pivotal campaign junctures, such as the conventions and the final

12 The invisible primary is "the long period of time after the previous general election but before the beginning of the next primary season" during which various potential candidates attempt to "gather support…raise funds, procure the support of other elected officials (endorsements) and party activists, court the favor of agents in the press, and generally increase their name recognition (Jackson 2002, 2, 5)."

13 The U.S. has experienced an enormous increase in early voting since 2000. As Kyle Yoder et al (2024) note, as of 2000, "40% of all voting-age citizens lived in states that offered at least one" early voting option. As of 2024, that number has increased to 97%.

14 In a highly unusual turn of events, during the 2024 cycle, the incumbent President, Joe Biden was initially the party's nominee with little contest. After he withdrew on July 21, 2024, the party turned to his Vice President, Kamala Harris, who secured the nomination and was formally anointed at the Democratic National Convention in Chicago in late-August.

Table 7.5: Cases of genAI use by timing.

Timing	Sub-Period	Cases	Percentage of Sub-Period	Cumulative % of Total
Pre-Election (PE)	Invisible Primary (Mar-Dec ‘23)	12	29%	20%
	Primaries (Jan-Jun ’24)	13	31%	22%
	Conventions (Jul-Aug ‘24)	17	40%	29%
PE Subtotal		42	100%	71%
Polling (PO)	Early Voting (Sep 1-Nov ‘24)	16	94%	27%
	Election Day (Nov 5 ’24)	1	6%	2%
PO Subtotal		17	100%	29%
Total		59	–	100%

weeks before Election Day. This knowledge is critical for campaigns, election authorities, and technology platforms aiming to allocate resources efficiently in countering GenAI-driven manipulation.

7.6 RQ4 – findings: GenAI misuse and beneficial use

Public apprehension about artificial intelligence (AI) in the electoral context reached considerable heights in the lead-up to the 2024 U.S. election. Multiple surveys reflected this anxiety, with a YouGov poll indicating that 85% of Americans were “very or somewhat concerned” about disinformation spread through deepfakes (Orth and Bialik 2023). A similarly high proportion emerged from Elon University data, which noted that seven out of ten respondents feared AI’s capacity to generate fabricated images, video, or audio (Elon University 2023). Pew Research findings revealed that 57% of Americans were “extremely or very concerned” about AI-driven misinformation, and another 25% expressed moderate concern, leading to over 80% acknowledging at least some level of unease (Gracia 2024).

Before the 2024 campaign cycle, real-world observations about these worries were limited. However, newly available empirical research on AI usage in elections—carried out worldwide during this period—offers fresh insight. Stockwell et al. (2024, 8–26) propose a conceptual split between “AI-enabled election threats” and “Beneficial AI election use cases.” Their categorization situates smear campaigns, deceptive advertising, and other disinformation attempts on one side, while highlighting potential advantages like fact-checking or innovative campaign practices on the other.

The current study follows a comparable dual framework, contrasting "detrimental (malicious) use" with "beneficial use." Here, detrimental usage refers to scenarios in which AI undercuts democratic norms by enabling disinformation or disruptive campaign practices. Beneficial usage entails applications that bolster the electoral process, for instance by broadening voter participation or reducing communication costs for under-resourced candidates. As with any classification, varying degrees of harmful or helpful effects exist, but the primary distinction lies in whether the AI deployment ultimately weakens or strengthens the broader democratic structure.

7.6.1 Misuse: Deepfakes, undisguised attacks, positive propaganda

As Table 7.6 shows, of the 59 identified instances of AI use, 45 (76%) involved misuse, while approximately one in five (21%) exemplified a beneficial application. One remaining case (2%) was categorized as "neutral," as it involved a purely satirical AI-produced video featuring Vice President Kamala Harris and Donald Trump in fictional romantic scenarios (Hilarious 2024). Although this neutral instance did not appear to undermine or enhance the democratic process, it still served as an example of how AI can create credible but whimsical content.

Table 7.6: Summarizes the distribution of misuse and beneficial uses.

Type of Use	Number of Cases	Percentage
Misuse (Malicious)	45	76%
Beneficial	12	21%
Neutral	1	2%
Total	58	99%

*One case's classification is missing due to incomplete data.

Although the media has often emphasized AI's role in spreading falsehoods deliberately, establishing creators' intentions can be challenging. Consequently, the three types of "GenAI misuse" highlighted in this study—deepfakes, undisguised attacks, and positive propaganda—are defined primarily by (a) how closely the content mimics an authentic source and (b) the tone of its message, ranging from negative to positive. In all three categories, the common element is that the AI-based material conveys some degree of falsehood, whether through outright fabrication or exaggeration.

Deepfakes typically strive for maximal realism, making them appear "authentic." Most are negative or neutral in tone, though positive deepfakes exist as well. The AI-enabled Biden robocall in New Hampshire was an especially prominent example, as

were several videos circulated by campaigns or private actors purporting to show public figures engaging in out-of-character behavior (Khan 2024; Seitz-Wald 2024; Nehamas 2023). Some deepfakes were more benign, involving fabricated endorsements by popular celebrities such as Elton John or Taylor Swift (Delzer 2024).

Table 7.7 presents a typology for GenAI misuse, showcasing three levels of realism—high, medium, and low—cross-referenced with the tone of the message (negative, neutral, or positive).

Table 7.7: GenAI misuse categories by degree of realism and message tone.

	High Realism	Medium Realism	Low Realism
Negative Tone	Deepfakes	Undisguised Attack	Undisguised Attack
Neutral Tone	Deepfakes	Undisguised Attack	Undisguised Attack
Positive Tone	Deepfakes	Positive Propaganda	Positive Propaganda

A further dimension involves "pink slime" or "hyper-local" fake news sites, occasionally attributed to foreign entities (Hall 2024; Watts 2024). These platforms mimic local media outlets yet consist entirely of fabricated AI-driven stories, tailor-made for distinct ideological or geographic audiences. While their overall quantity remains unclear, researchers suggest they represent a coordinated effort to disrupt public trust in legitimate information channels.

As shown in Table 7.8, deepfakes account for roughly six in ten cases (62%) of misuse, overshadowing two other categories: undisguised attacks (24%) and positive propaganda (13%). Deepfakes can occasionally be upbeat or apparently humorous (for example, fictional celebrity endorsements). Still, most documented deepfakes in the 2024 cycle were intended to tarnish reputations or convey misleading narratives about key figures in the election.

Table 7.8: Cases of genAI misuse by type.

Misuse Type	Number of Cases	Percentage
Deepfakes	28	62%
Undisguised Attacks of "Chaos"	11	24%
Total	45	99%*

*Rounding may cause the total to be 99%

Undisguised attacks differ from deepfakes by openly admitting that the material is AI-generated. They often focus on negative content—whether ridiculing opponents, provoking scandal, or sowing general confusion. One noteworthy example features an entirely AI-generated video from the Republican National Committee, depicting a dystopian scenario under another Biden presidency (Stepansky 2023). Another instance

was a controversial AI depiction of a Republican gubernatorial candidate in North Carolina, which included disclaimers that the images were manufactured (Yaede 2024). A borderline example emerged when an image circulated online showing Donald Trump supposedly in handcuffs; while it could pass for a deepfake, the creator publicly acknowledged it was produced with AI (Lajka 2023).

Positive propaganda likewise reveals its synthetic origins, but its content skews "fun" or entertaining to cast a candidate or viewpoint in a flattering light. These AI constructs often involve parodic dances, memes, or short clips that depict candidates in whimsical situations. A prominent example was a viral video of Elon Musk and Donald Trump dancing to the Bee Gees' "Stayin' Alive" (Robins-Early 2024), along with memes showing Trump performing spectacular feats with science fiction elements (Bond 2024). While these may appear less malicious, they still hinge on artificial illusions that can muddle public perception.

7.6.2 Beneficial use: Increasing engagement and assisting campaigns

This study located twelve cases (21%) of GenAI use geared toward strengthening democratic engagement or assisting campaign operations. According to Table 7.9, the vast majority (92%) involved helping campaigns manage standard tasks—such as fundraising, phone banks, or targeted ads—particularly benefiting under-resourced candidates who lack the financial means for large staffs.

One prominent example concerned Shamaine Daniels, a Pennsylvania Democrat who had previously fallen short against incumbent Republican Scott Perry. Ahead of a prospective 2024 rematch, Daniels partnered with "Ashley," an AI "volunteer" capable of personalizing telephone outreach to voters on a mass scale (Tong and Coster 2023). Similarly, the Democratic National Committee (DNC) experimented with AI to craft initial drafts of fundraising appeals, often outperforming purely human-written messages (Goldmacher 2023).

A second category—representing 8% of beneficial uses—entailed increasing voter awareness or encouraging participation. NextGen America, a youth-voter advocacy group, developed an AI chatbot called "Vote-E," which operated on the Discord platform to engage younger Black and Latino voters (Carrasquillo 2024). Though only a single instance of this type emerged in the identified dataset, it underscores how nonprofits or other civic organizations might employ AI to educate and mobilize otherwise disengaged communities.

Although the volume of beneficial applications is smaller than that of misuse, these examples reveal avenues through which AI can strengthen electoral processes: reaching marginalized communities, disseminating reliable information, and cutting operational costs for campaigns. In turn, greater use of such tools could potentially reduce the incentives for exclusively negative or deceptive strategies, giving research-

Table 7.9: Cases of beneficial genai use by type.

Beneficial Use Type	Number of Cases	Percentage
Increasing Voter Knowledge/Engagement	1	8%
Assisting Campaign Tasks (Under-Resourced)	11	92%
Total	12	100%

ers, policymakers, and civic organizations a clearer set of promising practices to encourage.

7.6.3 Source and timing as key inputs for risk strategy

While understanding the origin and timing of AI-driven activities is valuable, certain raw totals can be misleading. For instance, foreign state actors may comprise a lower share of overall cases but often prove more difficult and resource-intensive to counter because they operate across borders and require cooperation with multiple governments and platforms (Janjeva et al. 2023). Moreover, source and timing should not be treated in isolation. To establish robust risk strategies, it is vital to grasp whether foreign actors engage more heavily in one segment of the election cycle than domestic ones, or how different types of domestic sources vary in their activity levels over time.

Table 7.10 offers an example of such integrative data, correlating the timing of GenAI use with foreign vs. domestic origin.

Table 7.10: Cases of genai use by timing and source (foreign vs. domestic).

Timing	% Foreign	% Domestic
Pre-Election (PE)		
Invisible Primary (’23)	8%	26%
Primaries (Jan-Jun ’24)*	0%	28%
Conventions (Jul-Aug) 24**	25%	28%
Polling Period (PO)		
Early Voting (Sep 1-Nov 4, ’24)	67%	19%
Election Day (Nov 5, ’24)*	0%	0%

*=1 Source unknown; **=2 sources unknown.

As Table 7.10 shows, almost all (92%) foreign-sourced AI usage materialized between July and October 2024, spanning the conventions and early voting periods. Domestic usage was more evenly spread throughout a longer duration, beginning with the invisible primary in March 2023 and extending through the close of early voting.

Table 7.11 refines these patterns by subdividing foreign and domestic sources into state vs. non-state actors. It then maps their activities against the key election periods. Foreign state actors, for example, appear to concentrate on specific windows—mainly convention season and early voting—while domestic campaigns, political organizations, and individuals spread their efforts over a longer horizon.

Table 7.11: Cases of genai use by timing and actor type.

Timing	% Foreign	% Domestic
	% State	% Non-State
Pre-Election (PE)		
Invisible Primary (‘23)	0%	2%
Primaries (Jan-Jun ‘24)*	0%	0%
Conventions (Jul-Aug ’24)**	5%	0%
Polling Period (PO)		
Early Voting (Sep 1-Nov 4, ’24)	15%	0%
Election Day (Nov 5, ’24)*	0%	0%

These data underscore a need for flexible resource allocation. For foreign state actors who concentrate on the late summer and autumn period, risk-mitigation efforts might be intensified during these months. Domestic sources, by contrast, require monitoring from the earliest stages of the campaign, given their protracted engagement.

7.7 Conclusion

Public anxiety surrounding AI in elections frequently centers on the possibility that AI-enabled content might dramatically change or “steal” an election (Yan et al. 2025; Gracia 2024; Elon University 2023). Thus far, however, neither the present study nor others (Stockwell et al. 2024b) have identified definitive evidence of such a first-order risk.

Instead, emerging data point to considerable second-order risks, which from a democratic standpoint may be even more troubling. These include the deliberate use of AI by foreign and domestic actors to spread disinformation, sow confusion, and diminish public confidence in the authenticity of information. As Stockwell et al. (2024a, 3) note, deeper risks may involve eroding trust in legitimate content (including election results) or even inciting hatred or violence against public institutions. Since these second-order threats are more challenging to detect and control, they pose genuine concerns for long-term democratic stability.

7.7.1 Risk strategy development: RQs 1–3

I. RQ1: Presence of GenAI in the 2024 U.S. Election
As shown in Table 7.2 (see Section 4), this investigation documented 59 total instances of AI use. The emphasis on public media reporting in identifying these examples implies that the actual frequency of use may be higher, as sensational cases are more likely to be covered by traditional outlets and social networks.

II. RQ2: Source of GenAI Content
Analysis revealed that most attributable uses originated from domestic actors, while approximately one-fifth stemmed from foreign entities (Table 7.3). Of the international contributions, the majority were state-sponsored. Conversely, the domestic instances were generally linked to political organizations and campaigns (Table 7.4).

III. RQ3: Timing of AI Usage
Table 7.5 indicates that GenAI use was most prevalent during the pre-election window, peaking in July and August 2024, when the political conventions took place. Though one-quarter of all incidents happened in the polling period (early voting and Election Day), the pre-election phase (particularly after primary season) accounted for the largest cluster.

7.7.2 Risk mitigation strategy: RQs 4–5

IV. RQ4: Nature and Types of Use (Malicious vs. Beneficial)
Public discussions following the election often suggested that GenAI neither "materialized" as a major threat nor that deepfakes shaped outcomes significantly (Frenkel 2024; Wong 2024). While it is true no election was "stolen" through AI, analysis in this study revealed that nearly two-thirds of malicious GenAI incidents involved deepfakes (Table 7.8). Some of these were "positive" in tone, such as fabricated celebrity endorsements, but many were designed to mislead and to appear genuine. The more subtle distinction lies between well-disguised "deep" fakes and obviously exaggerated or clearly labeled "cheap" fakes.

Though malicious uses vastly outnumbered beneficial examples, approximately one-fifth of documented cases promoted democratic functions, such as boosting voter engagement or streamlining campaign tasks (Table 7.9). Future inquiry could clarify whether cost barriers or lack of training impede such positive applications.

V. RQ5: Toward a Comprehensive Risk Mitigation Strategy
Several implications emerge from these preliminary findings. First, continued inquiry into the who, when, and how of GenAI usage is essential, as a deeper evidence base is a prerequisite for any serious policy or regulatory intervention. Second, clarity in def-

initions matters greatly: terms like "deepfake" are used broadly in common language, but consistent categorization is vital for data-based decisions. Third, effective mitigation requires triangulating multiple factors—source, timing, type, motives, and the authenticity level of the AI content. This complexity extends to questions such as which organizations underwrite foreign state actors, or how disclaimers might shape audience interpretation of AI-generated material. Fourth, sampling issues remain: capturing only viral or media-driven cases may distort the real incidence and severity of GenAI usage.

Ultimately, even if no electoral outcome has been decisively overturned by AI, GenAI can still exercise significant influence through second-order effects. Important regulatory questions arise. Would disclaimers solve the problem of disinformation, or would they merely push fabricators to label content superficially? Might satirical or obviously faked images still undermine public trust? If so, how can societies preserve the basic credibility of political discourse?

From the vantage point of this "super-election year," three takeaways stand out. First, GenAI is likely to remain part of campaigns worldwide. Second, its impact is often subtle, working through doubt and confusion rather than blatant interference (second versus first-order effects). Third, methodically studying both the damaging and beneficial potentials of AI is crucial to designing robust risk strategies.

References

Artificial Intelligence (2024). *Oxford References*. February 9. https://www.oxfordreference.com/display/10.1093/oi/authority.20110803095426960

Barrabi, Thomas (2023). "AI deepfakes poised to wreak havoc on 2024 presidential election: experts," *New York Post*. June 14. https://nypost.com/2023/06/14/ai-deepfakes-poised-to-wreak-havoc-on-2024-election-experts/

Bond, Sheri (2024). "How AI deepfakes polluted elections in 2024." *All Things Considered/National Public Radio (NPR)*. December 21. https://www.npr.org/2024/12/21/nx-s1-5220301/deepfakes-memes-artificial-intelligence-elections

Bousquette, Isabelle (2024). "Underdog Who Beat Biden in American Samoa Used AI in Election Campaign," *Wall St. Journal*. March 8. https://www.wsj.com/articles/underdog-who-beat-biden-in-american-samoa-used-ai-in-election-campaign-b0ce62d6

Brack, Sebastien (2024). "AI In Elections: Comparing the EU and US experiences in 2024," *Kofi Annan Foundation*. December 18. https://www.kofiannanfoundation.org/news/ai-in-elections-comparing-the-eu-and-us-experiences-in-2024/

Carrasquillo, Adrian (2024). "Democrats use AI in effort to stay ahead with Latino and Black voters," *The Guardian*. August 21. https://www.theguardian.com/us-news/article/2024/aug/21/democrats-ai-black-latino-voters

Delzer, Erielle and Kara Fellows (2024). "Can you spot the AI impostor? The Swift rise of AI celebrity endorsements," *CBS News*. October 16. https://www.cbsnews.com/news/rise-of-ai-celebrity-endorsements/

Denkovski, Ognjan and Daniela Alvarado Rincón (2024). "The Generative AI Factor at the Ballot Box: A review of Generative AI Use in the 2024 European Parliament Elections," *Kofi Annan Foundation in*

collaboration with Democracy Reporting International. October. file:///Users/jeannezaino/Downloads/The-GenAI-Factor-at-the-Ballot-Box_compresses.pdf

DeWitt, Ethan (2024). "Biden wins the New Hampshire Democratic primary, without showing up," *New Hampshire Bulletin*. January 23. https://newhampshirebulletin.com/2024/01/23/biden-wins-the-new-hampshire-democratic-primary-without-showing-up/

Elon University News Bureau Staff (2023). "New survey finds that most Americans expect AI abuses will effect 2024 election," *Today at Elon*. May 15. https://www.elon.edu/u/news/2024/05/15/ai-and-politics-survey/

Escott, Eban (2017). What are Three Types of AI: A Guide to Narrow, General, and Super Artificial Intelligence," Codebots. October 24. https://codebots.com/artificial-intelligence/the-3-types-of-ai-is-the-third-even-possible

"Fact Check: Video does not show Joe Biden making transphobic remarks" (2023). *Reuters*. February 10. https://www.reuters.com/article/fact-check/video-does-not-show-joe-biden-making-transphobic-remarks-idUSL1N34Q1IW/

"Fact Check: Image altered to include AI-generated Kamala Harris in McDonald's uniform," (2024). *Reuters*. November 4. https://www.reuters.com/fact-check/image-altered-include-ai-generated-kamala-harris-mcdonalds-uniform-2024-10-29/

"FCC Proposes $6M Fine for Robocalls" (2024). *News Release*, *Federal Communications Commissions (FCC)*, May 23, https://www.fcc.gov/document/fcc-proposes-6m-fine-nh-robocalls

Frenkel, Sheera (2024). "The Year of the A.I. Election That Wasn't," New York Times. August 21. https://www.nytimes.com/2024/08/21/technology/ai-election-campaigns.html

Goldmacher, Shane (2023). "A Campaign Aide Didn't Write That Email. A.I. Did," *New York Times*. March 28. https://www.nytimes.com/2023/03/28/us/politics/artificial-intelligence-2024-campaigns.html

Goldstein, Josh and Andrew Lohn (2024). "Deepfakes, Elections, and Shrinking the Liars Dividend," *Brennan Center for Justice*. January 23. https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend

Gracia, Shanay (2024). "Americans in both parties are concerned over the impact of AI on the 2024 presidential campaign," *Pew Research Center*. September 19. https://www.pewresearch.org/short-reads/2024/09/19/concern-over-the-impact-of-ai-on-2024-presidential-campaign/

Hall, Sheri (2024). "AI and the Election," *The Link*. Fall. https://magazine.cs.cmu.edu/ai-and-the-election

"Hilarious AI-Generated video shows Donald Trump and Kamala Harris in romantic scenarios" (2024). *MundoDeportivo*. August 18. https://www.mundodeportivo.com/us/en/20240818/697970/hilarious-ai-generated-video-shows-donald-trump-and-kamala-harris-in-romantic-scenarios.html

Hsu, Tiffany and Cade Metz (2024). "In Big Election Year, A.I.'s Architects Move Against Its Misuse," *New York Times*. February 16. https://www.nytimes.com/2024/02/16/technology/ai-elections-defense.html

Jackson, Joseph C (2002). "Party Animal: The Frontrunner in the Presidential Invisible Primary," The University of Georgia. August. https://getd.libs.uga.edu/pdfs/jackson_joseph_c_200208_ma.pdf

Janjeva, Ardi, Alexander Harris, Sarah Mercer, Alexander Kasprzyk and Anna Gausen (2023). "The Rapid Rise of Generative AI Assessing risks to safety and security," *The Centre for Emerging Technology and Security (CETaS), The Alan Turing Institute*. December. https://cetas.turing.ac.uk/sites/default/files/2023-12/cetas_research_report_-_the_rapid_rise_of_generative_ai_-_2023.pdf

Johnson, Derek B (2024). "FBI flags false videos impersonating agency, claiming Democratic ballot fraud," *Cyberscoop*. November 2. https://cyberscoop.com/fbi-fake-videos-ballot-fraud-democrats-doppelganger/

Johnson, Jonathan (2024). "Types of AI with Examples," *bmc blogs*. November 11. https://www.bmc.com/blogs/artificial-intelligence-types/

Khan, Thalia (2024). "What Role Did A.I. Play In the 2024 Election," *Partnership on AI*. November 4. https://partnershiponai.org/what-role-did-ai-play-in-the-2024-u-s-election/

Klepper, David (2024). “Russia is Behind Viral Disinformation Targeting Walz, Intelligence Official Says,” *Associated Press (AP)*. October 22. https://apnews.com/article/walz-russia-disinformation-election-harris-trump-8646b7f354df629cd5d6661ba51e7174

Kochi, Sudiksha (2023). “Fact check: Photos showing Trump arrested by law enforcement are computer-generated,” *USA Today*. March 22. https://www.usatoday.com/story/news/factcheck/2023/03/22/fact-check-photos-trump-arrest-created-artificial-intelligence-ai/11520957002/

Lajka, Arijeta. 2023. “Fake AI images of Putin, Trump being arrested spread online,” PBS News. March 23. https://www.pbs.org/newshour/politics/fake-ai-images-of-putin-trump-being-arrested-spread-online

Le Monde (2024). “Macron Calls Snap Elections After Historic Far Right Gains in European Vote.” *Le Monde/AFP*. June 9. https://www.lemonde.fr/en/international/article/2024/06/09/macron-calls-early-elections-after-historic-far-right-gains-in-european-vote_6674311_5.html

Luca de Tena, Carlos (2024). “67% of Europeans fear AI manipulation in elections, according to IE University research,” *Center for the Governance of Change/IE University*. October 21. https://www.ie.edu/university/news-events/news/67-europeans-fear-ai-manipulation-elections-according-ie-university-research/

“Meta says AI had only ‘modest’ impact on global elections in 2024,” (2024). *Al Jazeera*. December 3. https://www.aljazeera.com/news/2024/12/3/meta-says-ai-had-only-modest-impact-on-global-elections-in-2024

Nehamas, Nicholas (2023). DeSantis Campaign Uses Apparently Fake Images to Attack Trump on Twitter, *New York Times*. June 8. https://www.nytimes.com/2023/06/08/us/politics/desantis-deepfakes-trump-fauci.html

Orth, Taylor and Carl Bialik (2023). “Majorities of Americans are concerned about the spread of AI deepfakes and propaganda,” *YouGov*. September 12. https://today.yougov.com/technology/articles/46058-majorities-americans-are-concerned-about-spread-ai

Paxton, Ben, Rebecca McKee, and Jack Pannell (2024). “General Election 2024: What Happens When a General Election is Called?” May 24. *Institute for Government*. https://www.instituteforgovernment.org.uk/explainer/uk-general-election-july-2024

Rehagen, Tony (2024). “Candidate AI: The Impact of Artificial Intelligence on Elections.” *Emory Magazine*. Fall. https://news.emory.edu/features/2024/09/emag_ai_elections_25-09-2024/index.html

Robins-Early, Nick (2024). “Trump posts deepfakes of Swift, Harris and Musk in effort to shore up support,” *The Guardian*. August 19. https://www.theguardian.com/us-news/article/2024/aug/19/trump-ai-swift-harris-musk-deepfake-images

Saab, Beatriz (2024). “Manufacturing Deceit – How Generative AI Supercharges Information Manipulation”, *National Endowment for Democracy*. June 18. file:///Users/jeannezaino/Downloads/The-GenAI-Factor-at-the-Ballot-Box_compresses.pdf

Schneier, Bruce and Nathan Sanders (2024). “The apocalypse that wasn’t: AI was everywhere in 2024’s elections, but deepfakes and misinformation were only part of the picture.” *The Conversation*. December 2. https://theconversation.com/the-apocalypse-that-wasnt-ai-was-everywhere-in-2024s-elections-but-deepfakes-and-misinformation-were-only-part-of-the-picture-244225

Seitz-Wald, Alex and Mike Memoli (2024). “Fake Joe Biden robocall tells New Hampshire Democrats not to vote Tuesday,” *NBC News*. January 22. https://www.nbcnews.com/politics/2024-election/fake-joe-biden-robocall-tells-new-hampshire-democrats-not-vote-tuesday-rcna134984

Sheehan, Jeanne (2021). *American Democracy in Crisis: The Case for Rethinking Madisonian Government*. New York: Palgrave Macmillan.

Seemela, Ontlotlile (2023). “Positive Propaganda: A Double-Edged Necessary Tool,” *ArtsHelp*. https://www.artshelp.com/positive-propaganda/

Steffen, Sarah (2024). “AI-driven bot network tried to help Trump win US election,” DW. November 5. https://www.dw.com/en/2024-us-election-ai-driven-social-media-bots-donald-trump-supporters/a-70695526

Stepansky, Joseph (2023). "Wild West: Republican Video Shows AI Future in US Elections," *Al Jazeera*. April 28. https://www.aljazeera.com/news/2023/4/28/wild-west-republican-video-shows-ai-future-in-us-elections

Stockwell, Sam, Megan Hughes, Phil Swatton and Katie Bishop (2024a). "AI-Enabled Influence Operations: The Threat to the UK General Election." *The Centre for Emerging Technology and Security (CETaS), The Alan Turing Institute*. May. https://cetas.turing.ac.uk/sites/default/files/2024-05/cetas_briefing_paper_-_ai-enabled_influence_operations_-_the_threat_to_the_uk_general_election.pdf

Stockwell, Sam (2024). "AI-Enabled Influence Operations: Threat Analysis of the 2024 UK and European Elections," The Threat to the UK General Election." *The Centre for Emerging Technology and Security (CETaS), The Alan Turing Institute*. September. https://cetas.turing.ac.uk/sites/default/files/2024-09/cetas_briefing_paper_-_ai-enabled_influence_operations_-_threat_analysis_of_the_2024_uk_and_european_elections.pdf

Stockwell, Sam, Megan Hughes, Phil Swatton, Albert Zhang, Jonathan Hall KC, and Kieran (2024b). "AI-Enabled Influence Operations: Safeguarding Future Elections," *The Centre for Emerging Technology and Security (CETaS), The Alan Turing Institute*. November. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-safeguarding-future-elections

Swenson, Ali and Will Weissert (2024). "AI robocalls impersonate President Biden in an apparent attempt to suppress votes in New Hampshire," *Associated Press/PBS News*. January 22. https://www.pbs.org/newshour/politics/ai-robocalls-impersonate-president-biden-in-an-apparent-attempt-to-suppress-votes-in-new-hampshire

"The 2024 Global Elections Super-Cycle" (2024). *International IDEA*, December 15. https://www.idea.int/initiatives/the-2024-global-elections-supercycle

Thomas, Elise (2024). "I've found a large," *X*, November 3. https://threadreaderapp.com/thread/1853208626565206502.html

Tong, Anna and Helen Coster (2023). Meet Ashley, the world's first AI-powered political campaign caller," *Reuters*. December 15. https://www.reuters.com/technology/meet-ashley-worlds-first-ai-powered-political-campaign-caller-2023-12-12/

Ulmer, Alexandra and Anna Tong (2023). "Deepfaking It: America's 2024 Election Collides with the AI Boom," *Reuters*. May 30. https://www.reuters.com/world/us/deepfaking-it-americas-2024-election-collides-with-ai-boom-2023-05-30/#:~:text=Welcome%20to%20America's%202024%20presidential,polarized%20world%20of%20U.S.%20politics.

Watts, Clint (2024). "Iran Targeting 2024 US Election," *Microsoft*. August 8. https://blogs.microsoft.com/on-the-issues/2024/08/08/iran-targeting-2024-us-election/

Wong, Matteo (2024). "AI's Fingerprints Were All Over the Election," *The Atlantic*. November 15. https://www.theatlantic.com/technology/archive/2024/11/ai-election-propaganda/680677/

Yaede, Caitlyn (2024). "'Over the top' ad features a fake Mark Robinson. What to know about AI in political ads," *The Charlotte Observer*, *Yahoo!*. October 10. https://www.pressreader.com/usa/charlotte-observer/20241016/281659670499725?srsltid=AfmBOorIAodUardEWMIYVXLc8N9-MsUkK78SPxFfVfdZmF0PU1Csi2mL

Yaojun Yan, Harry, Garrett Morrow, Kei-Chang Yang, and John Wihbey (2025). "The origin of public concerns over AI supercharging misinformation in the 2024 U.S. presidential election," *Misinformation Review*. January 30. https://misinforeview.hks.harvard.edu/article/the-origin-of-public-concerns-over-ai-supercharging-misinformation-in-the-2024-u-s-presidential-election/

Yoder, Kyle, April Tan, and Stefan Martinez-Ruiz (2024). "The Expansion of Voting Before Election Day, 2000-2024." *The Center for Election Innovation & Research*. July. https://electioninnovation.org/research/expansion-voting-before-election-day/

Zaino, Jeanne Sheehan (2024). "Watching the U.K. Election Gives a Feeling of Electoral Envy," *The Fulcrum*. July 17. https://thefulcrum.us/money-politics/uk-election

C. Christopher Lee and Heechang Shin

8 Preserving privacy in healthcare: Implications of artificial intelligence

8.1 Introduction

Artificial intelligence (AI) refers to the ability of computers to perform tasks typically carried out by humans (Khalid et al., 2023). However, for the AI to work properly, it requires large volumes of data to train the model and function effectively. Therefore, it is essential to provide the vast volume of data for training AI models. In healthcare, AI is bringing new possibilities. For example, AlphaFold, a recent AI-powered protein structure prediction algorithm solved the protein folding problem that hampered crucial advancements in biology and medicine for the past 50 years (Jumper et al., 2021). The benefits of using AI could include rapid diagnosis, personalizing care, and reducing unnecessary outpatient appointments that could save billions to the economy (Lee and Yoon, 2021).

Vast amounts of data for training AI models raises significant concerns about data privacy—especially when that data includes sensitive and confidential patient information (Khalid et al., 2023). Medical or health privacy refers to the protection of patient records and the confidentiality of communications among healthcare professionals (Khalid et al., 2023). Patients often worry about their health information being shared with insurance companies, employers, or other third parties, as it can be used in ways that negatively impact them. For instance, sharing medical history with insurers may lead to discrimination, denial of coverage, or increased premiums. To safeguard individual privacy rights, it is essential to implement robust protections and ensure responsible data handling by third parties (Khalid et al., 2023).

In response to such concerns, many countries have enacted legislation to protect personal privacy. In the United States, the Health Insurance Portability and Accountability Act (HIPAA), passed in 1996, was designed to enhance legal protections for healthcare facilities and secure patient information (Terry, 2007). Also, strict regulations such as the European Union's General Data Protection Regulation (GDPR) mandate rigorous compliance when working with healthcare datasets (Khalid et al., 2023). However, this approach does not ensure confidentiality from the providers themselves, and patients have no way to independently verify whether providers are following their stated privacy policies, making this approach insufficient to fully meet privacy and security requirements (Haas et al., 2011).

To address the privacy and security concerns surrounding sensitive patient data, there is a pressing need for the healthcare community to enhance data governance frameworks and develop privacy-preserving approaches that enable AI innovation without compromising patient confidentiality. Any breach of healthcare data sources

https://doi.org/10.1515/9783111595306-008

or AI models—allowing an adversary to access private information such as patient identities, health records, or locations—is highly undesirable and represents a serious threat to user privacy. Qayyum et al. (2020) emphasize the importance of addressing data privacy and security concerns when handling healthcare data. Similarly, Hall and McGraw (2014) warn that both providers and patients may lose trust in telehealth solutions if the underlying technologies lack robust security and privacy safeguards. Tom et al. (2020) highlight the critical need for data protection in the context of AI-enabled ophthalmology. Additionally, Mamdouh et al. (2020) draw attention to privacy challenges associated with the Internet of Things (IoT), focusing on its potential impact on healthcare privacy.

Some may argue that encrypting medical records is sufficient to protect health privacy. However, this approach does not secure data if both patient information and the encryption/decryption keys are stored in the same server. If a data breach occurs that exposes both the encrypted data and the keys, the patient data can still be accessed. Even when data and keys are stored separately, risks remain. Developing AI models typically requires direct access to raw data, and as Khalid et al. (2023) note, sensitive patient information must be transmitted to the compute server for model training. This transmission itself is a point of vulnerability, potentially exposing data to interception. Moreover, the AI model can pose significant privacy risks. When trained on sensitive patient data, the AI models can unintentionally memorize and leak identifiable information through their outputs. This allows adversaries to reconstruct training data or determine whether a specific individual's data was used during training. Even without direct access to raw data, the model's parameters can encode patterns that reveal private details. As AI systems grow more powerful and widely accessible, the risk of such unintended disclosures increases, making the entire system highly susceptible to both insider threats and external attacks. For a comprehensive overview of these attacks, see the work by (Haas et al., 2011).

This chapter presents a survey of the literature on protecting data privacy in the development of healthcare AI systems. Section 2 provides an overview of AI applications in the healthcare industry and discusses the associated privacy and security concerns. Section 3 examines potential solutions to address these challenges, focusing on privacy-preserving techniques for AI model development. Finally, Section 4 concludes with a summary of key insights and future directions.

8.2 Use of AI in healthcare industry

AI is transforming the healthcare industry by improving diagnostics, patient monitoring, healthcare management, and drug discovery. This section reviews these key benefits while also examining the privacy and security challenges associated with the use of AI in healthcare.

8.2.1 Improving diagnostics

AI is used to help improving diagnostic accuracy. Donnelly et al. (2024) developed an AI system called AsymMirai which uses deep learning techniques to detect breast cancer up to five years in advance. Donnelly et al. (2024) used the publicly accessible EMory BrEast imaging dataset which includes a range of lesion annotations, pathology results, and demographic data for more than 116,000 patients. The performance of AsymMirai varied across different racial groups, indicating the need for future efforts to address and potentially reduce the risk of perpetuating or worsening health care disparities. MIT researchers have developed "Mirai," an AI model that detects breast cancer (Yala et al., 2021). The model predicts precancerous changes with high accuracy, factoring in age and family history and can identify invasive cancer cells in tissue that looks normal, enabling earlier, more effective treatment and improving survival rates, particularly for high-risk women (Yala et al., 2021). Wu et al. (2024) used AI for cancer detection and diagnosis. Deep learning algorithms were used cytology and cervical neoplastic disease detection.

AI models are also used for predicting diseases. Fitriyani et al. (2020) utilized machine learning techniques such as DBSCAN, SMOTE-ENN, and XGBoost to diagnose the subjects' heart disease status earlier. Using Statlog and Cleveland datasets, the model achieved 95.90% and 98.40% accuracy outperforming other models such as Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest. A prototype model was developed to assist clinicians in diagnosing heart disease, enabling timely intervention and reducing mortality risks.

8.2.2 Patient monitoring

Patient monitoring has become one of the main technologies for delivering proactive preventive care to patients (Sneha and Varshney, 2009). Wearable devices continuously track physiological parameters which is used for patient monitoring (Pal et al., 2023). These devices enables healthcare providers to take proactive measures, preventing complications, and AI-enabled remote monitoring systems manage conditions such as hypertension and chronic obstructive pulmonary disease (COPD), where timely treatment adjustments can lead to significantly improved patient outcomes (Jatawan et al., 2023). AI-powered electrocardiograms embedded in wearable devices have demonstrated high accuracy in detecting arrhythmias, enabling early interventions and reducing emergency hospital admissions (Siontis et al., 2021). Similarly, AI-enhanced continuous glucose monitors assist diabetes patients by predicting blood sugar fluctuations and recommending insulin adjustments, enhancing glycemic control (Huang et al., 2025).

8.2.3 Health management

IBM Watson for Oncology uses AI to analyze medical literature, patient records, and clinical trial data to recommend personalized cancer treatment plans (Zhou et al., 2019). In gastroenterology, AI is used for diagnosing and treating diseases by detecting and classifying polyps and personalizing treatment plans using patient data, leading to better outcomes and reduced side effects. (Rammohan et al., 2024). Participants expressed strong interest and perceived value as long as it served as a supplementary tool rather than replacing physician judgment (Zhou et al., 2019). Jain and Prajapati (2021) presented that AI enhances emergency decision-making, clinical predictions, early disease detection, and hospitalization forecasting, enabling proactive patient care. The availability of extensive clinical data has accelerated the adoption of deep learning, facilitating faster and more accurate patient data analysis, ultimately improving healthcare outcomes and operational efficiency (Jain and Prajapati, 2021)

8.2.4 Drug discovery

AI is revolutionizing drug discovery and personalized medicine. Qureshi et al. (2023) discussed that AI applications permeate various stages of the drug discovery pipeline, including de novo drug design, property prediction, molecular simulations, and candidate prioritization, and open-source databases and AI-driven tools facilitate drug design, though challenges persist in molecule representation, data quality, and labeling inconsistencies. Mak et al. (2024) provided a comprehensive overview of AI in drug discovery and development, detailing its potential, methodologies—including machine learning (ML), deep learning (DL), and natural language processing (NLP)—and real-world applications. Involvement of AI can aid rational drug design, assist in decision making, determine the right therapy for a patient, including personalized medicines, and manage the clinical data generated and use it for future drug development (Mak et al., 2024).

AI enhances patient selection, dosage determination, and monitoring, increasing the success rates of clinical trials (Anuyah et al., 2024). Integrating artificial intelligence (AI) into clinical trials has transformed drug development and personalized medicine by employing deep learning techniques—such as convolutional neural networks (CNNs) and transformer-based models—for patient stratification, adverse event prediction, and treatment personalization; predictive modeling approaches, including survival analysis and time-series forecasting, to enhance trial efficiency and reduce failure rates; and natural language processing (NLP) to extract actionable insights from unstructured clinical data, collectively streamlining workflows, improving patient outcomes, and reducing costs associated with trial inefficiencies (Anuyah et al., 2024).

8.2.5 Privacy challenges in the use of AI in healthcare

The introduction of machine learning techniques and AI make privacy a pressing issue in healthcare industry. Medical privacy or health privacy refers to the right of individuals to keep their medical information confidential and control who has access to it (Alpert, 2003), which includes personal health records, diagnoses, treatments, and any other sensitive health-related data. Terry and Francis (2007) discussed that while progress has been made toward implementing electronic health records, prioritizing patient and physician trust is required for success by ensuring strong privacy protections, granting patients full control over their personal information beyond the minimal safeguards of HIPAA, even if it means compromising some interoperability and system comprehensiveness.

Haas et al. (2011) presented the following four requirements for privacy-preserving electronic health records:

1) Patient Control Over Privacy Policies: Patients must have the ability to set binding privacy rules on how their health data is accessed, used, and disclosed.
2) Verification of Privacy Enforcement: Patients should be able to verify that their privacy policies are being followed and identify any unauthorized access or data leaks.
3) Limited Trust in Third Parties: Patients should only need to trust their direct healthcare providers and certified authorities. (Haas et al., 2011) stated that, in conjunction with Requirement 1, this requirement ensures that the provider cannot access patient-related data beyond what is necessary for the service or modify medical records stored in the system.
4) Prevention of Data Linkage for Profiling: Combining different medical data flows should not allow unauthorized profiling or the discovery of new private information about patients.

The first requirement, patient control over privacy policies, emphasizes user-defined access control and ensuring patient data privacy. However, ensuring this requirement can negatively impact model performance, especially when sensitive data are excluded from training (Kaissis et al., 2020). When patients restrict certain data, those features become unavailable to AI models, even if they are highly predictive, which limits model's prediction performance. The second requirement, verification of privacy enforcement, introduces additional complexity in model development. Engineers cannot assume access of all the dataset or easily aggregate large datasets, forcing them to design models that can function with incomplete or missing inputs (Haas et al., 2011). Enforcing these privacy requirements while maintaining strong model performance is technically challenging and often demands advanced privacy-preserving techniques—such as federated learning or differential privacy—which may lower accuracy or increase system complexity (Haas et al., 2011). As a result, balancing privacy protection with the need for high-quality predictive models remains a difficult and ongoing challenge in health-

care AI. The next section examines potential solutions to address these privacy challenges.

8.3 Privacy preserving techniques

8.3.1 Cryptography-based approach

Homomorphic Encryption

Homomorphic encryption is a cryptographic method that allows arithmetic operations to be performed directly on encrypted data without the need for decryption. This property makes it a powerful tool for securely outsourcing data analysis, as it enables encrypted data to be processed without exposing the underlying sensitive information (Acar et al., 2018). For instance, RSA supports multiplicative homomorphism, allowing encrypted numbers to be multiplied without decryption. However, RSA is considered a Partially Homomorphic Encryption (PHE) scheme because it supports only one operation—multiplication. In contrast, the Paillier cryptosystem supports additive homomorphism, enabling the addition of encrypted numbers without decryption (Acar et al., 2018).

Building on these foundations, Lu et al. (2015) proposed a homomorphic encryption-based approach for secure, outsourced analysis of large-scale genomic data. Carpov et al. (2016) developed a privacy-preserving solution that supports health data analysis in the cloud without revealing personal information. Their method processes encrypted frequency tables derived from genomic and clinical data to calculate key statistical metrics such as Hardy-Weinberg Equilibrium (which checks for population genetic balance), the Chi-Square test (to assess associations between variables), and Linkage Disequilibrium (to measure non-random associations of genetic variants). To improve computational efficiency, they used a packing technique, which combines multiple data values into a single ciphertext, reducing processing time.

More recently, Sarkar et al. (2021) investigated privacy-preserving genotype imputation—predicting missing genetic information—by combining machine learning models with homomorphic encryption to develop logistic regression models and deep learning, integrating non-linear functions to enhance prediction accuracy. To support secure exponential calculations within encrypted environments, they employed Taylor series approximations and implemented the computations (Chillotti et al., 2019).

Homomorphic encryption satisfies the four requirements discussed in section 2.5. Because the data never needs to be decrypted for processing, patients retain greater control over what is shared and how it's used. While the protocol itself does not directly provide full auditability, when used with cryptographic logging (Moghaddam, 2022), patients and auditors can verify that data processing adhered to the agreed-

upon policies without seeing the underlying data. Since third parties (e.g., cloud services or analytics providers) only see encrypted data and cannot decrypt it, patients do not have to fully trust these parties. Only the patient or a certified authority holding the decryption key can decrypt the results. By keeping data encrypted during all operations, the scheme prevents unauthorized correlation or linkage between datasets that could lead to profiling. Since entities processing the data never see the plaintext, they cannot combine different sources or infer private information across datasets.

Secure Multiparty Computation

Secure Multi-party Computation (SMC) is a cryptography technique that enables multiple parties to jointly compute a function over their inputs while keeping those inputs private (Bayatbabolghani and Blanton, 2018). Each party applying the algorithm to its secure data without knowing the rest of the data results in privacy preservation. (Khalid et al., 2023). Therefore, multiple data owners can combine their private inputs to compute a mathematical function without revealing their inputs to other people, and in this way, each data owner can have ownership of their data. For example, the average salary of a group can be calculated without revealing any individual's salary by using a secure multi-party computation technique called additive secret sharing (Bayatbabolghani and Blanton, 2018). In the healthcare domain, this capability is particularly valuable for protecting sensitive patient information. Akgün et al. (2022) introduced an efficient, privacy-preserving method for querying large-scale genomic databases using two non-colluding proxies to outsource encrypted data, maintaining data, query, and output privacy in semi-honest cloud environments. Similarly, Gong et al. (2016) addressed privacy concerns in genome-aware health monitoring systems by proposing an SMC-based approach that enables users to receive diagnostic results derived from genomic and biomedical sensing data, without revealing any information to the healthcare provider. Barni et al. (2011) developed a privacy-preserving automatic diagnosis system, allowing remote servers to classify biomedical signals without accessing either the input data or the final diagnostic results. Li et al. (2020) proposed a self-service medical diagnosis scheme that combines SMC and homomorphic encryption, allowing patients to obtain accurate diagnoses online while protecting both their medical data and hospital diagnostic models. In response to the COVID-19 pandemic, Kumar et al. (2020) introduced a privacy-preserving self-care health management model using MPC, enabling secure online sharing of encrypted patient data for remote diagnosis while preventing unauthorized access. Additionally, Marwan et al. (2016) designed a cloud-based collaborative healthcare framework using SMC, facilitating secure data sharing and computation among healthcare organizations without compromising individual patient privacy.

Secure multi-party computation can be used to satisfy the four requirements discussed in section 2.5. Patients retain control over who can participate in the computa-

tion and under what terms, and the protocol can be designed to respect granular, patient-defined privacy rules, ensuring that functions can be computed and results disclosed privately. The protocol enforces patient control by preventing unauthorized data access during computation, ensuring data is only used for explicitly approved purposes. While the protocol itself does not directly provide audit logs, it can be combined with cryptographic proofs such as zero-knowledge proofs to allow verification that the computation was performed correctly and according to policy without leaking any private information. Since the data remains hidden throughout, and the computation is mathematically constrained by the protocol, unauthorized access or deviation can be detected or provably prevented. The protocol reduces reliance on third parties by keeping data distributed and encrypted throughout computation, requiring trust only in the secure execution of the protocol, and the protocol protects against profiling by never exposing raw data, even when datasets are combined for joint analysis.

8.3.2 Non-cryptographic techniques

Differential Privacy

Differential privacy is a mathematical framework for protecting privacy when sharing series of data. It allows organizations to share useful information about a dataset while ensuring that the privacy of any individual in the dataset is not compromised (Dwork, 2006). Differential privacy ensures that the output of a data analysis is essentially the same whether any one individual's data is included in the dataset, which means that no attacker can confidently determine whether a particular person's data was used, even if they have access to other information (Dwork, 2006). To achieve this, random noise is added to the results of computations on the data.

Abadi et al. (2016) presented methods to enhance differentially private training by tracking higher moments of privacy loss for tighter estimates, improving computational efficiency with new gradient and batching techniques, and demonstrating practical implementation in deep learning, showing that strong privacy can be achieved with minimal impact on performance and complexity. Deep learning is an AI technique that works by combining many simple layers, like basic math operations and activation functions to turn inputs into outputs. By adjusting the settings of these layers, the network is trained to match a set of example inputs and their correct outputs. The goal of training is to find the best set of parameters that make the average loss across all examples as small as possible, ideally minimizing it as much as we can. Their implementation and experiments demonstrated that deep neural network was trained with non-convex objectives, under a modest privacy budget, and at a manageable cost in software complexity, training efficiency, and model quality (Abadi et al.,

2016). The model starts with the randomly chosen starting values, and for each round of training, the following steps are performed:

1) Pick a random subset of data from the full dataset.
2) For each example selected in the previous step, calculate how the model should change to reduce the error, called a gradient.
3) Limit the size of each gradient so that no single data point can have too much influence (this is called clipping the gradient).
4) Take the average of these clipped gradients.
5) Add some random noise to this average. The noise helps hide the impact of any one person's data, which is key for protecting privacy.
6) Update the model parameters by taking a small step in the direction that reduces the error, based on the noisy average gradient.

The above processes are repeated, improving the model while keeping each person's data private. After training is done, you get the final model and calculate how much privacy was "used up" during training, which tracks the overall privacy cost (Abadi et al., 2016).

Differential privacy satisfies the four privacy needs for electronic health records, based on (Haas et al., 2011). First, differential privacy helps protect patient control over their personal health data by making sure that individual information cannot be revealed, even when data is shared or analyzed. Although it doesn't give patients direct tools to set privacy rules, it automatically ensures that no single person's data can be singled out in reports or research. This means that patient data is protected by design, and their private health information stays hidden regardless of how the data is used.

Second, differential privacy makes it possible to verify that privacy is being enforced. It does this by using something called a "privacy budget" that tracks how much private information could be revealed through repeated queries (Ebadi et al., 2015). A privacy budget is a way to control how much private information can be revealed from a dataset. When someone analyzes data that contains personal details, each question or analysis can potentially reveal a small amount about individuals in the data. To protect privacy, systems that use differential privacy add a bit of random noise to the answers. This noise hides the exact details but still allows for useful insights. The privacy budget keeps track of how much privacy has been "used up" through these queries. Each time you ask a question; you spend a little bit of the budget. If you ask too many questions, or if the answers are too precise, the budget runs out. Once it's gone, the system won't give more detailed answers to avoid risking anyone's privacy. It's a balance — the more accurate the results, the more of the privacy budget you use. This way, the system allows you to learn general trends from data without exposing any one person's private information. Systems using differential privacy can log and measure this budget, giving both patients and providers a way to

confirm that no one's privacy has been violated. This helps build trust and allows for clear auditing of how the data was accessed or used.

Third, one of the key benefits of differential privacy is that it reduces the need to trust the third parties involved. Because the data is protected even after being shared or analyzed, patients only need to trust their doctor or healthcare provider—not everyone who might work with the data. Researchers or analysts can perform useful studies without ever seeing private patient details.

Finally, differential privacy prevents people from combining different data sources to learn private information or build hidden profiles of patients. It adds controlled noise to data so that results are accurate on aggregate level but don't reveal anything specific about individuals. This makes it much harder for someone to connect data points and discover something private, protecting patients from profiling or unwanted data linkages.

Federated Learning

Federated learning is a way to train machine learning models without centralizing all the data on one server. Rather than collecting and storing data in a single location, the training is distributed across multiple devices or local systems—such as user phones or hospital databases—where the data already resides. The model is sent to these locations to learn directly from the local data. After training, only the model updates (not the raw data) are sent back to a central server.

The primary advantage of federated learning is enhanced privacy. Since raw data stays on the user's device or within the institution, sensitive information remains secure and local. This is particularly important in fields like healthcare, where protecting patient data is critical. By transmitting only model parameters or weight updates—not the actual data—federated learning helps prevent individual information from being exposed or stored outside its original source.

In addition to privacy, federated learning is efficient. It eliminates the need to transfer large datasets across networks, reducing communication overhead and bandwidth consumption. Its decentralized nature also makes it highly scalable, ideal for settings where data is distributed across numerous sources, such as smartphones, IoT devices, or large organizations.

In real-world applications, federated learning is being used in areas like medical diagnostics in hospitals. For example, multiple hospitals can collaborate to train a model that identifies disease patterns without ever sharing patient records. This combination of collaboration and privacy makes federated learning a powerful tool for advancing data-driven insights while respecting data ownership and confidentiality. Dou et al. (2021) showed that federated learning can effectively train a privacy-preserving AI model to detect COVID-19-related CT abnormalities, generalizing well across multinational datasets without sharing sensitive patient data. Qayyum et al.

(2022) proposed privacy-sensitive, low-latency COVID-19 diagnosis using federated learning, which showed up to 16% improvements on X-ray and ultrasound data over centralized models. Key challenges and deployment considerations for edge-based ML in healthcare were also discussed. Roth et al. (2020) demonstrated the effectiveness of federated learning for breast density classification across seven global clinical institutions. Without sharing data, FL models outperformed local models by 6.3% on average and showed a 45.8% improvement in generalizability across sites.

This section reviews the work of Lee et al. (2018), who proposed a privacy-preserving platform for federated patient similarity learning across institutions. In their framework, patient-level data remains local, yet the system can identify clinically similar patients across hospitals. The goal of patient similarity learning is to find individuals with comparable clinical profiles within specific medical contexts, such as heart disease or cancer. This federated approach supports real-world applications like building diverse patient cohorts, tracking disease outbreaks, and recruiting for clinical trials. The process involves two key steps: learning context-specific patient representations from electronic health records (EHRs) and using these representations to find similar patients. Implementing this in a federated setting poses challenges, including creating effective and adaptable representations, managing heterogeneous data, and ensuring privacy throughout the entire pipeline.

The proposed approach involves three key components. First is feature construction, where patient health records are transformed into numerical formats suitable for computational analysis. For example, gender may be encoded as 0 or 1, while diagnosis codes, medications, and lab results are translated into vectors indicating the presence or frequency of specific events. This step produces a detailed digital representation of each patient's medical history.

Second is hashing, which compresses each patient's digital profile into a binary string. This transformation not only enables efficient similarity comparison but also enhances privacy. Given the diverse nature of patient data—ranging from demographics to lab results—the model employs a multi-hash approach, generating separate hash codes for each type of data. This allows the system to more accurately capture the similarities across different dimensions of patient information.

Finally, the federated patient matching framework ensures that all patient data remains at its original site. Hospitals independently compute hash codes using shared hash functions, and only these codes are used in cross-site similarity searches. When a query is made, the system compares these hashed representations to identify matches without exposing raw data. The similarity is calculated using mathematical operations on the hash codes, and the model is trained using known patient relationships (e.g., patients diagnosed with the same disease). The framework includes optimization techniques to efficiently train the model, avoid overfitting, and ensure stable performance across distributed institutions.

Although hash codes offer greater privacy than raw data, they can still be vulnerable to reverse engineering (Lee et al., 2019). To mitigate this risk, the authors pro-

posed a hybrid approach, a federated learning with incorporating homomorphic encryption. In the proposed scheme, a hospital encrypts its query patient's data and sends it to another hospital. That hospital then performs similarity comparisons between the encrypted query and its own encrypted patient data, returning the result in encrypted form. Only the querying hospital can decrypt the result to view the similarity score. At no point is any hospital exposed to another's raw or even encrypted patient data. The only information exchanged is a mathematically protected similarity score, such as a distance metric, ensuring that privacy is preserved throughout the process. This method enables secure, privacy-preserving patient matching across hospitals and regions.

Federated learning supports the goals of privacy-preserving electronic health records and aligns well with the four key requirements proposed by Haas et al. (2011). By design, federated learning keeps data localized—at hospitals, clinics, or individual devices—while still enabling collaborative model training. This decentralized approach offers several advantages for privacy and control.

First, federated learning helps support patient control over privacy policies by ensuring that raw health data never leaves the institution or device where it originates. Instead of sending data to a central server, only model updates—such as weight adjustments—are shared. This approach respects patient consent by keeping control of sensitive data with the patient's healthcare provider, which can enforce specific privacy rules aligned with each patient's preferences.

Second, federated learning can enable verification of privacy enforcement through techniques like audit logging and privacy-preserving model tracking. Since data access does not occur directly, and only model updates are transmitted, the opportunities for unauthorized access are reduced. Additionally, federated learning frameworks can integrate differential privacy or secure multiparty computation to further strengthen transparency and offer verifiable privacy guarantees, helping detect or prevent data misuse.

Third, federated learning directly supports limited trust in third parties. Because patient data remains with the original provider and is not pooled on a central server, patients only need to trust their healthcare institutions, not any external data processors or aggregators. This aligns closely with the requirement that patients should not have to trust entities beyond their direct care providers or certified authorities. Federated learning minimizes data exposure and access, which reduces the risk of unauthorized changes or misuse.

Finally, federated learning helps prevent data linkage for profiling by design. Since data is not centralized, there is no single database where an attacker could combine records from multiple sources to build detailed patient profiles. Even if an adversary were to intercept model updates, these updates are typically aggregated and do not contain identifiable information. Additional tools like differential privacy can be layered into federated learning systems to obscure patterns that could otherwise be used for patient profiling.

In summary, federated learning aligns strongly with all four privacy requirements identified by (Haas et al., 2011), offering a robust framework for secure and privacy-preserving health data analysis across multiple institutions.

8.4 Conclusion

Although AI holds tremendous potential to revolutionize healthcare through improving diagnostics, patient monitoring, healthcare management, and drug discovery, its reliance on large volumes of sensitive data introduces privacy concerns. Addressing these concerns is critical not only for regulatory compliance but also for maintaining trust among patients and providers. This chapter has examined current challenges and threats associated with healthcare data sharing and AI development, as well as practical approaches to safeguard privacy. As AI continues to shape the future of healthcare, developing robust, privacy-preserving systems will be essential to ensure ethical and responsible innovation.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security* (pp. 308–318).

Acar, A., Aksu, H., Uluagac, A. S., & Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)*, *51*(4), 1–35.

Alpert, S. A. (2003). Protecting medical privacy: challenges in the age of genetic information. *Journal of Social Issues*, *59*(2), 301–322.

Anuyah, S., Singh, M. K., & Nyavor, H. (2024). Advancing clinical trial outcomes using deep learning and predictive modelling: Bridging precision medicine and patient-centered care. *arXiv preprint arXiv:2412.07050*.

Akgün, M., Pfeifer, N., & Kohlbacher, O. (2022). Efficient privacy-preserving whole-genome variant queries. *Bioinformatics*, *38*(8), 2202–2210.

Barni, M., Failla, P., Lazzeretti, R., Sadeghi, A. R., & Schneider, T. (2011). Privacy-preserving ECG classification with branching programs and neural networks. *IEEE Transactions on Information Forensics and Security*, *6*(2), 452–468.

Bayatbabolghani, F., & Blanton, M. (2018). Secure multi-party computation. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security* (pp. 2157–2159).

Chillotti, I., Gama, N., Georgieva, M., & Izabachène, M. (2019). TFHE: Fast fully homomorphic encryption library.

Carpov, S., Nguyen, T. H., Sirdey, R., Constantino, G., & Martinelli, F. (2016, June). Practical privacy-preserving medical diagnosis using homomorphic encryption. In *2016 ieee 9th international conference on cloud computing (cloud)* (pp. 593–599). IEEE.

Donnelly, J., Moffett, L., Barnett, A. J., Trivedi, H., Schwartz, F., Lo, J., & Rudin, C. (2024). AsymMirai: interpretable mammography-based deep learning model for 1–5-year breast cancer risk prediction. *Radiology*, *310*(3), e232780.

Dugan, T., & Zou, X. (2016). A survey of secure multiparty computation protocols for privacy preserving genetic tests. In *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)* (pp. 173–182). IEEE.

Dou, Q., So, T. Y., Jiang, M., Liu, Q., Vardhanabhuti, V., Kaissis, G., . . . & Heng, P. A. (2021). Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. *NPJ digital medicine*, *4*(1), 60.

Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming* (pp. 1–12). Berlin, Heidelberg: Springer Berlin Heidelberg.

Ebadi, H., Sands, D., & Schneider, G. (2015). Differential privacy: Now it's getting personal. *Acm Sigplan Notices*, *50*(1), 69–81.

Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2020). HDPM: an effective heart disease prediction model for a clinical decision support system. *Ieee Access*, *8*, 133034–133050.

Gong, Y., Zhang, C., Hu, Y., & Fang, Y. (2016). Privacy-preserving genome-aware remote health monitoring. In *2016 IEEE Global Communications Conference (GLOBECOM)* (pp. 1–6). IEEE.

Hall, J. L., & McGraw, D. (2014). For telehealth to succeed, privacy and security risks must be identified and addressed. *Health Affairs*, *33*(2), 216–221.

Haas, S., Wohlgemuth, S., Echizen, I., Sonehara, N., & Müller, G. (2011). Aspects of privacy for electronic health records. International journal of medical informatics, 80(2), 26–31.

Huang, W., Pang, I., Bai, J., Cui, B., Qi, X., & Zhang, S. (2025). Artificial Intelligence-Enhanced, Closed-Loop Wearable Systems Toward Next-Generation Diabetes Management. *Advanced Intelligent Systems*, 2400822.

Jain, K., & Prajapati, V. (2021). NLP/deep learning techniques in healthcare for decision making. *Primary Health Care: Open Access*, *11*(3), 373–380.

Jatawan, V., Jaiswal, C., & Khandelwal, R. (2023). Optimizing Nanodrug Formulation And Delivery Using Artificial Intelligence: A Cutting-Edge Approach For Personalized Medicine. Advancements In Immunology, Nanomedicine. *Advancements In Immunology, Nanomedicine, And Public Health*, *63*.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., . . . & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *nature*, *596*(7873), 583–589.

Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, *2*(6), 305–311.

Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine*, *158*, 106848.

Lee, D., & Yoon, S. N. (2021). Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. *International journal of environmental research and public health*, *18*(1), 271.

Li, D., Liao, X., Xiang, T., Wu, J., & Le, J. (2020). Privacy-preserving self-serviced medical diagnosis scheme based on secure multi-party computation. *Computers & Security*, *90*, 101701

Lu, W. J., Yamada, Y., & Sakuma, J. (2015). Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption. In *BMC medical informatics and decision making* (Vol. 15, pp. 1–8). BioMed Central.

Kumar, A. V., Sujith, M. S., Sai, K. T., Rajesh, G., & Yashwanth, D. J. S. (2020). Secure Multiparty computation enabled E-Healthcare system with Homomorphic encryption. In *IOP Conference Series: Materials Science and Engineering* (Vol. 981, No. 2, p. 022079). IOP Publishing.

Lee, J., Sun, J., Wang, F., Wang, S., Jun, C. H., & Jiang, X. (2018). Privacy-preserving patient similarity learning in a federated environment: development and analysis. *JMIR medical informatics*, *6*(2), e7744.

Mak, K. K., Wong, Y. H., & Pichika, M. R. (2024). Artificial intelligence in drug discovery and development. *Drug discovery and evaluation: safety and pharmacokinetic assays*, 1461–1498.

Mamdouh, M., Awad, A. I., Hamed, H. F., & Khalaf, A. A. (2020). Outlook on security and privacy in IoHT: Key challenges and future vision. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)* (pp. 721–730). Springer International Publishing.

Marwan, M., Kartit, A., & Ouahmane, H. (2016). Applying secure multi-party computation to improve collaboration in healthcare cloud. In 2016 third international conference on systems of collaboration (SysCo) (pp. 1–6). IEEE.

Moghaddam, P. (2022). *A framework for secure logging and analytics in precision healthcare cloud-based services* (Doctoral dissertation).

Pal, S., Kumari, K., Kadam, S., & Saha, A. (2023). The ai revolution. *IARA Publication*.

Qayyum, A., Ahmad, K., Ahsan, M. A., Al-Fuqaha, A., & Qadir, J. (2022). Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge. *IEEE Open Journal of the Computer Society*, *3*, 172–184.

Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2020). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, *14*, 156–180.

Qureshi, R., Irfan, M., Gondal, T. M., Khan, S., Wu, J., Hadi, M. U., . . . & Alam, T. (2023). AI in drug discovery and its clinical relevance. *Heliyon*, *9*(7).

Rammohan, R., Joy, M. V., Magam, S. G., Natt, D., Magam, S. R., Pannikodu, L., . . . & Mustacchia, P. (2024). Understanding the landscape: the emergence of artificial intelligence (AI), ChatGPT, and Google Bard in gastroenterology. *Cureus*, *16*(1).

Rashid, M., & Sharma, M. (2025). AI-Assisted Diagnosis and Treatment Planning—A Discussion of How AI Can Assist Healthcare Professionals in Making More Accurate Diagnoses and Treatment Plans for Diseases. *AI in Disease Detection: Advancements and Applications*, 313–336.

Roth, H. R., Chang, K., Singh, P., Neumark, N., Li, W., Gupta, V., . . . & Kalpathy-Cramer, J. (2020). Federated learning for breast density classification: A real-world implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020*, *Lima, Peru*, *October 4–8*, *2020*, *Proceedings 2* (pp. 181–191). Springer International Publishing.

Sarkar, E., Chielle, E., Gürsoy, G., Mazonka, O., Gerstein, M., & Maniatakos, M. (2021). Fast and scalable private genotype imputation using machine learning and partially homomorphic encryption. *IEEE access*, *9*, 93097–93110.

Sneha, S., & Varshney, U. (2009). Enabling ubiquitous patient monitoring: Model, decision protocols, opportunities and challenges. *Decision Support Systems*, *46*(3), 606–619.

Siontis, K. C., Noseworthy, P. A., Attia, Z. I., & Friedman, P. A. (2021). Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, *18*(7), 465–478.

Terry, N. P., & Francis, L. P. (2007). Ensuring the privacy and confidentiality of electronic health records. *U. Ill. L. Rev*., 681.

Tom, E., Keane, P. A., Blazes, M., Pasquale, L. R., Chiang, M. F., Lee, A. Y., . . . & Force, A. A. I. T. (2020). Protecting data privacy in the age of AI-enabled ophthalmology. *Translational vision science & technology*, *9*(2), 36–36.

Wu, T., Lucas, E., Zhao, F., Basu, P., & Qiao, Y. (2024). Artificial intelligence strengthens cervical cancer screening–present and future. *Cancer Biology & Medicine*, *21*(10), 864–879.

Yala, A., Mikhael, P. G., Strand, F., Lin, G., Smith, K., Wan, Y. L., ... & Barzilay, R. (2021). Toward robust mammography-based models for breast cancer risk. Science Translational Medicine, 13(578).

Zhou, N., Zhang, C. T., Lv, H. Y., Hao, C. X., Li, T. J., Zhu, J. J., . . . & Zhang, X. C. (2019). Concordance study between IBM Watson for oncology and clinical practice for patients with cancer in China. *The oncologist*, *24*(6), 812–819.

Yusuf Joseph Ugras

9 Artificial intelligence in accounting

9.1 Introduction

Since 1980's, the accounting profession has undergone a significant technological transformation that has streamlined processes, increased accuracy, improved compliance, and provided deeper insights into financial data. This evolution of technology has profoundly transformed the accounting profession, yielding numerous benefits beyond enhanced efficiency, and accuracy, and brought increased strategic value to those that rely on accounting information. Traditionally, accountants maintained physical ledgers, recording every transaction by hand. This method was labor-intensive, time-consuming, and prone to errors. The introduction of computerized accounting systems in the late 20th century replaced manual ledgers with software that quickly processes transactions, reduces human error, and increases speed. Programs like QuickBooks and Peachtree revolutionized how financial data was recorded and managed. Another technological innovation came through spreadsheets. Before spreadsheets, accountants relied heavily on paper-based computations, which limited their ability to perform complex calculations and scenario analyses. The development of spreadsheet software to work on personal computers started with VisiCalc and Lotus 123, and spreadsheets have become dominant tool with the advanced features of Microsoft Excel. The software allows accountants to automate calculations with use of macros, easily manipulate data, and perform complex financial forecasts with power queries. Functions like "What-If Analysis" empowered accountants to evaluate different scenarios and optimize business decisions. AI implementations to spreadsheets are now taking place. As businesses grew, so did the complexity of their financial operations. Manual methods couldn't keep up with the growing volume of transactions. Accounting software solutions streamlined bookkeeping processes, providing features like invoicing, payroll management, and tax filing. The accounting software not only saved time but also improved the detail and accuracy of financial reporting. In recent years, cloud computing transformed accounting which was previously dominated by desktop software, making collaboration and accessibility challenging, especially for remote teams. Cloud-based accounting systems provide anytime, anywhere access to financial data. They facilitate real-time collaboration among teams, enhance client communications, and allow for automatic software updates and backup. This transition to the cloud has also reduced the need for extensive IT infrastructure.

https://doi.org/10.1515/9783111595306-009

9.2 Risks of technology in accounting

The rapid advancement of technology discussed above has significantly transformed accounting, offering increased efficiency, accuracy, and scalability. However, these innovations also introduce several risks that can impact businesses, organizations, and individuals. The implementation and use of computerized accounting systems, spreadsheets, and cloud-based accounting present various challenges that require careful management. While many businesses have adapted to these changes and established countermeasures to address potential threats, it is crucial to critically assess the risks associated with each technology.

9.2.1 Computerized accounting systems (CAS)

Computerized accounting systems have become integral to modern financial management, yet they pose substantial risks that organizations must mitigate. One of the primary concerns involves security risks. Since CAS entails the digital storage and processing of financial data, there is an elevated risk of unauthorized access and data breaches. Cyberattacks, such as hacking or phishing, could expose sensitive financial information, leading to severe financial and reputational consequences. Furthermore, insufficient encryption methods for data transfer and storage increase the vulnerability of accounting data to theft or tampering.

Beyond security threats, system failures and downtime present operational challenges. Technical malfunctions, software bugs, or hardware failures can delay financial reporting or result in the loss of crucial financial records. Additionally, software updates or newly implemented applications may be incompatible with existing systems, potentially leading to errors in data transfer, calculations, or reporting.

Another risk is fraud, particularly internal fraud. Employees or system users with access to specific accounting modules may exploit vulnerabilities to manipulate financial records. If automated fraud detection mechanisms are not correctly programmed, fraudulent activities may go undetected.

Finally, an over-reliance on technology can lead to a decline in manual oversight. Accountants who excessively depend on automated systems may neglect essential manual verification procedures, increasing the likelihood of undetected discrepancies or errors that computerized systems fail to flag.

9.2.2 Spreadsheets (e.g., Microsoft excel)

While spreadsheets are widely used in financial analysis and reporting, they are susceptible to several risks. One of the most significant concerns is human error in data entry and formula application. Since spreadsheets rely heavily on user-inputted for-

mulas, small mistakes—such as a misplaced decimal point or an incorrectly referenced cell—can lead to substantial errors in calculations. Manual data entry increases the likelihood of inconsistencies, which can result in incorrect financial reporting.

Additionally, version control issues are common in spreadsheet-based workflows. In collaborative environments, multiple versions of the same file may circulate among users, leading to confusion, redundancy, and errors. Unlike specialized accounting software, spreadsheets often lack a comprehensive audit trail, making it difficult to track changes and identify who made specific modifications.

Another limitation of spreadsheets is their inability to scale effectively. While they function well for smaller datasets, they become inefficient when handling large volumes of transactions. Excessively large spreadsheets can be slow, prone to crashing, and susceptible to data loss, limiting their usefulness for complex financial tasks.

In addition, security vulnerabilities present a serious risk. Sensitive financial data stored in spreadsheets may be shared via unsecured channels such as email or external drives, increasing the risk of unauthorized access. Furthermore, spreadsheets typically lack sophisticated access controls, allowing multiple users unrestricted access to confidential financial information.

9.2.3 Accounting software (e.g., quickbooks, xero, sage)

The adoption of accounting software has streamlined financial operations; however, it introduces several implementation and operational risks.

One major challenge is implementation and integration. Transitioning from manual or legacy accounting systems to modern software can be complex and time-consuming. If the data migration process is not executed properly, businesses risk data loss or inconsistencies. Furthermore, accounting software may not seamlessly integrate with other business management systems, such as customer relationship management (CRM) or enterprise resource planning (ERP) platforms, leading to inefficiencies and potential data synchronization issues.

Security risks also accompany accounting software usage. Despite built-in security features, vulnerabilities in the software or improper configurations can lead to unauthorized access or data corruption. Additionally, when businesses rely on third-party vendors for technical support and maintenance, they expose themselves to the risk of external access to sensitive financial data.

Another concern is vendor dependency. Businesses that rely heavily on a particular software provider may face challenges if the vendor discontinues a product, changes pricing structures, or goes out of business. Subscription-based models, which are common for modern accounting platforms, also pose financial risks if unexpected price increases strain company budgets.

Finally, user training and adaptation are essential for ensuring efficient software use. Many accounting software platforms have a learning curve, and inadequate

training can lead to improper usage, errors in financial reporting, and underutilization of available features.

9.2.4 Cloud computing

Cloud computing has revolutionized financial management by offering remote access to accounting systems and data storage solutions. However, it also introduces significant data privacy, accessibility, and cybersecurity risks.

A key concern is data privacy and compliance. Many jurisdictions impose strict regulations on data storage and processing, such as the General Data Protection Regulation (GDPR) in Europe. If cloud services fail to comply with these regulations, businesses may face legal consequences and financial penalties. Furthermore, cloud data centers are often located in different jurisdictions, leading to uncertainty about data ownership, privacy, and control.

Another major risk relates to access and availability. Cloud-based accounting systems depend on a stable internet connection, meaning any disruption in connectivity could impede access to critical financial data. Additionally, service outages and downtime—whether due to technical failures or scheduled maintenance—can disrupt business operations, particularly for organizations that rely on real-time financial reporting.

Cybersecurity threats remain a persistent challenge for cloud-based accounting solutions. While cloud providers implement robust security measures, they are frequent targets of cyberattacks. Threats such as Distributed Denial of Service (DDoS) attacks or ransomware incidents can compromise financial systems, leading to data breaches or operational disruptions. Additionally, multi-tenant cloud environments introduce the risk of data being accessed by unauthorized parties if security protocols are not effectively enforced.

Lastly, vendor lock-in can limit business flexibility. Migrating financial data and processes between cloud providers can be complex, costly, and time-consuming. Businesses that become dependent on a specific provider may find it difficult to switch due to proprietary technologies, contractual obligations, or high migration costs.

Technological advancements in accounting and finance have revolutionized business operations by enhancing efficiency, accuracy, and scalability. However, these innovations also introduce significant risks related to data security, fraud, system failures, and regulatory compliance. Organizations must proactively address these risks through robust cybersecurity measures, regular training for financial professionals, and adherence to regulatory requirements. Furthermore, careful selection of technology vendors and ongoing system updates are crucial to mitigating potential vulnerabilities. By understanding and addressing these challenges, businesses can effectively leverage technology while safeguarding the integrity of their financial operations.

9.3 Artificial intelligence

On August 31, 1955 four academicians invited participants to a summer session at Dartmouth College to conduct research on what they called “Artificial Intelligence”, thereby starting the term AI (McCarthy et al. 2006) According to Davenport et al. (2019), Artificial Intelligence (AI) refers to technologies that simulate human intelligence and are designed to imitate human thinking and actions. Britannica explains that AI is the capacity of digital computers or computer robots to carry out tasks that typically require human intelligence. In US, the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence government defines AI as a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action (The White House, 2023).

While the discussion of definitions is of interest and meaningful, the organization must understand the benefits as well as the risks of AI use in the organization and have proper governance, risk management, and control tools. As technology continues to evolve with AI, accountants have the opportunity to use AI software to streamline routine processes and further enhance their own roles as strategic advisors, leveraging technology to create value for businesses and adapt to an increasingly complex financial landscape. As highlighted by Lawrence (2023), accountants can shift their focus towards more valuable financial activities, positioning themselves as key decision-makers and leaders within their firms. Embracing and understanding the technological advancements from AI will be crucial for accounting professionals looking to remain relevant and competitive in the future.

9.3.1 Artificial intelligence in accounting

The integration of Artificial Intelligence (AI) into accounting has significantly transformed the financial management landscape, providing notable benefits in automation, data analysis, fraud detection, cost efficiency, regulatory compliance, and strategic decision-making. As businesses increasingly adopt AI-driven solutions, accounting functions are becoming more efficient, precise, and integrated with broader business operations. The next sections discuss the various AI benefits to accounting.

9.3.1.1 Automation of repetitive tasks

One of the most immediate and tangible impacts of AI in accounting is the automation of repetitive tasks that traditionally are labor-intensive. Activities such as data entry, invoice processing, and reconciliations—traditionally time-consuming and prone to human error—are efficiently managed by AI-powered systems. By automating these routine processes, accountants can take advantage of these cost-saving opportunities, increase accuracies and shift their focus to more strategic, value-added activities that require human judgment and expertise. Data Entry tasks related to invoices, receipts, and other financial documents, could be eliminated reducing manual workload and minimizing errors. Bookkeeping Tasks of categorization of transactions, account reconciliations, and financial report generation could be handled through AI, streamlining bookkeeping activities. Traditional invoicing processes were often inefficient, involving physical documentation, mail delays, and manual payment processing. AI-powered e-invoicing and digital payment solutions have modernized financial transactions, reducing paperwork and improving accuracy. These technologies allow businesses to process payments more efficiently, enhance accounts receivable tracking, and facilitate quicker financial transactions, improving cash flow and operational effectiveness.

9.3.1.2 Enhanced data analysis and decision-making

Traditional accounting relied on rule-based systems with limited predictive capabilities. AI and machine learning (ML) have transformed this landscape, enabling accountants to analyze vast amounts of data and generate forward-looking insights. AI-powered analytics tools process vast amounts of financial data in real-time, identifying trends, patterns, and anomalies that might be overlooked through traditional methods. Unlike manual data analysis, which requires building tables and graphs to discern patterns, AI automates the data analytics process, rapidly analyzing large datasets, revealing insights that might not be immediately evident through a manual review. AI algorithms can use a data set involving prior sets of revenue data, integrated with various additional factors to perform a data analytics forecast of revenues.

With AI's ability to process and interpret real time data at an unprecedented speed, organizations can make more informed, data-driven financial decisions, improving forecasting accuracy, risk assessment, and overall business performance.

9.3.1.3 Fraud detection and security enhancements

Financial fraud remains a significant threat to businesses, but AI is proving to be an invaluable tool in mitigating this risk. AI algorithms can swiftly detect irregularities

in financial transactions and flag potentially fraudulent activities. By continuously learning from new data, AI adapts to increasingly sophisticated fraud attempts, strengthening financial security. AI-powered systems can identify suspicious transactions and anomalies, flagging them for further investigation and detecting fraud. In addition, AI continuously evolves and learns from emerging fraud patterns by making it a robust defense against financial crime. By reducing financial fraud risks, AI enhances the credibility of accounting practices, ensuring the protection of businesses and their stakeholders.

9.3.1.4 Optimized resource allocation

AI-driven forecasting and budgeting tools help businesses optimize resource allocation by analyzing historical data and identifying patterns. These insights enable companies to predict future financial and cash flow needs more accurately, allowing for strategic resource distribution. AI algorithms analyze past financial data, market trends, and other factors to improve budget planning. For firms with cash flow volatility, AI assists in monitoring and optimizing cash flow, ensuring financial stability and operational liquidity. Through predictive analytics, AI can enhance financial forecasting capabilities, providing businesses with data-backed predictions for revenue and expenditure. By optimizing financial resources, businesses can enhance efficiency, reduce unnecessary expenditures, and gain a competitive advantage in the market.

9.3.1.5 Enhanced analytics through data visualization

Traditional accounting reports were often static and provided limited analytical insights. Recently, data visualization software such as Tableau and Power BI have revolutionized financial reporting, making data more accessible and actionable. These software have started to integrate AI into their capabilities. Through this integration, AI enables accountants to use the interactive data and generate dynamic financial reports. The role of the accountants becomes more of a strategic advisor than simply a report generator role. By leveraging AI-driven visualization, accountants can present financial data in more engaging and comprehensible formats, improving stakeholder engagement and decision-making.

The integration of AI into accounting has revolutionized financial management, offering automation, enhanced data analysis, fraud detection, cost efficiency, and strategic forecasting. AI-driven tools empower accountants to focus on higher-value activities, strengthen financial security, and optimize business operations. As AI technology continues to evolve, businesses that embrace these innovations will be well-positioned to improve financial efficiency, mitigate risks, and drive long-term growth.

9.3.2 Negative consequences of AI

While the potential benefits of artificial intelligence (AI) in accounting and finance are substantial, they must be carefully weighed against the possible negative consequences. These risks can be categorized into three main areas: job displacement concerns, data security risks, and overreliance on AI.

One major concern is the potential displacement of human workers due to AI automation. According to the World Economic Forum's Future of Jobs Survey 2023, data entry clerks, administrative and executive secretaries, as well as accounting, bookkeeping, and payroll clerks are among the professions most at risk of decline over the next five years (World Economic Forum, 2023). As AI systems increasingly take over tasks that were traditionally performed by accountants and finance professionals, there is a growing need for reskilling and redeployment of human resources. However, according to the *Harvard Business Review*, AI is primarily intended to augment human capabilities rather than replace jobs entirely. This perspective suggests that AI should be viewed as a tool for enhancing efficiency and decision-making rather than as a threat to employment, AI Won't Replace Humans — But Humans With AI Will Replace Humans Without AI. (Lakhani 2023)

Another critical issue is data security. AI systems depend heavily on data, and any vulnerabilities in data storage or processing may expose sensitive financial information to security breaches. To mitigate this risk, organizations must implement robust security measures and regularly update their AI systems to protect against potential cyber threats. Ensuring the security of financial data is particularly essential for Chief Financial Officers (CFOs), who must take proactive steps to safeguard their organization's information.

Furthermore, excessive reliance on AI can lead to complacency and a decline in critical thinking among finance professionals. An overdependence on AI-generated insights without human oversight may result in errors or misinterpretations that could have significant financial consequences. To prevent this, it is crucial to maintain a balance between automation and human expertise. Organizations should ensure that AI-driven analyses are continually validated by professionals who can assess their accuracy and reasonableness.

While AI presents numerous opportunities for innovation in accounting, its risks must be carefully managed. Addressing concerns related to job displacement, data security, and overreliance on AI is essential to maximizing its potential while maintaining ethical and operational integrity.

9.3.3 AI in accounting firms

Large public accounting firms typically have two primary practices: Assurance services (audit of company financial report), and Tax services. Both areas are being im-

pacted by AI. In the next sections, the opportunities and risks of AI for these practices are explored.

9.3.3.1 Audit

The integration of artificial intelligence (AI) into audit practices offers several advantages that enhance efficiency, accuracy, and decision-making capabilities. AI-driven audit tools streamline repetitive tasks such as data extraction, reconciliation, and compliance checks, thereby allowing auditors to focus on high-value activities such as advisory services, forensic analysis, and comprehensive financial evaluations. Traditional auditing involves manually reviewing large volumes of financial data, a process that is both time-consuming and prone to human error. AI-driven auditing software can rapidly analyze financial statements, detect patterns, and identify discrepancies with unparalleled speed and precision. Through automating these labor-intensive procedures, AI reduces audit time and workforce expenditure. This cost efficiency benefits both accounting firms and their clients, making high-quality audits more accessible. The automation of audit procedures also allows firms to shift their focus from routine compliance tasks to more strategic and analytical engagements. Specifically, ChatGPT's capabilities in identifying potential non-compliance areas or reporting discrepancies are vital to accounting professionals (Kokina et al., 2025; Kokina & Davenport, 2017; Munoko et al., 2020, Zhao & Wang 2024).

Additionally, AI improves fraud detection and risk management by employing pattern recognition and predictive analytics to identify suspicious transactions. By analyzing historical data, AI can detect anomalies indicative of fraudulent activities, thereby reducing financial irregularities and minimizing audit risk. AI-driven audits improve risk assessment by leveraging machine learning algorithms to predict potential irregularities. Predictive analytics enable auditors to preemptively address risks rather than merely reacting to anomalies post-factum. This shift from a retrospective to a proactive audit approach marks a significant advancement in audit methodologies.

Moreover, AI contributes to strengthened compliance and regulatory oversight by ensuring adherence to evolving tax laws and financial reporting standards. Automated detection of discrepancies facilitates regulatory compliance, with oversight bodies increasingly recognizing AI's role in auditing. AI also supports advanced data analytics and decision-making by enabling real-time data analysis. This capability allows auditors to generate precise financial forecasts, identify inconsistencies in financial statements, and provide deeper insights into business risks. Consequently, firms can enhance their decision-making processes and conduct more effective, data-driven audits.

9.3.3.1.1 Risks in audit

Despite its numerous benefits, AI implementation in audit practices presents several challenges, including cybersecurity threats, regulatory compliance issues, and transparency concerns. One of the primary risks is fraud and cybersecurity vulnerabilities. Although AI enhances fraud detection, it can also be exploited by malicious actors who use AI to generate fraudulent financial documents or manipulate audit data, thereby increasing cybersecurity threats. As AI technologies evolve, so do the potential risks associated with their misuse.

Regulatory and compliance challenges also arise with AI integration in auditing. AI-generated financial reports must adhere to established auditing standards and legal frameworks. However, the absence of universally accepted guidelines for AI-driven audits complicates compliance efforts for firms. Another significant concern is the lack of transparency and explainability in AI models. Many AI algorithms function as "black boxes," meaning their decision-making processes are not easily interpretable by human auditors. This opacity raises credibility concerns regarding AI-generated audit findings and affects the accountability of firms that rely on such technologies.

Additionally, data privacy and confidentiality risks pose challenges, as AI systems require access to sensitive financial and client data. This increases the likelihood of data breaches and unauthorized access. Accounting firms must implement stringent data security measures to mitigate these risks. Lastly, professional liability risks arise when AI-generated errors lead to inaccurate financial reporting, potentially exposing firms to legal liabilities and reputational damage. Given that AI models rely on historical data, any biases or inaccuracies in the training datasets can negatively influence audit outcomes, resulting in flawed financial reports.

9.3.3.1.2 Managing AI risks in audit

To maximize the benefits of AI while mitigating associated risks, public accounting firms should implement robust AI risk management strategies. First, firms should establish comprehensive AI governance frameworks that outline ethical usage, regulatory compliance, and risk management protocols. These frameworks should include oversight mechanisms to ensure responsible AI deployment.

Second, firms must enhance data security and privacy measures by implementing strong encryption protocols, data access controls, and AI monitoring mechanisms. Clear policies on data sharing and AI access restrictions should be established to safeguard sensitive financial information. Additionally, ensuring transparency in AI models is crucial. Firms should adopt AI tools that provide auditability and explainability, allowing human auditors to validate AI-generated findings and maintain the integrity of the audit process.

Furthermore, continuous monitoring and human oversight are essential. AI should complement rather than replace human auditors. Regular human review and validation of AI-generated reports are necessary to ensure accuracy and reliability.

Establishing oversight teams to evaluate AI outputs for inconsistencies and biases can further enhance audit credibility. Lastly, accounting professionals should undergo comprehensive AI training and awareness programs to develop proficiency in AI applications, limitations, and ethical considerations in audit practices. By understanding AI's strengths and weaknesses, auditors can effectively leverage AI tools while maintaining professional skepticism.

Industry leaders in public accounting should also focus on developing standardized regulatory frameworks for AI-driven auditing to solidify AI's role in the audit profession. By adopting these recommendations, firms can harness AI's potential while mitigating risks, thereby ensuring more efficient, reliable, and compliant audit practices.

9.3.3.2 Tax

Artificial intelligence (AI) is transforming tax practices by enhancing automation, accuracy, and efficiency in tax compliance and advisory services. One of its most significant applications is tax research and compliance, where AI-powered tools, including generative AI models, automate research, identify applicable regulations, and cross-reference tax codes. These tools provide real-time updates on tax law changes, ensuring firms remain compliant.

AI also streamlines tax preparation and filing by extracting data from financial records, organizing deductions, and minimizing human errors. AI-driven software optimizes tax positions by analyzing historical tax data and detecting potential deductions. Additionally, machine learning algorithms enhance fraud detection and risk management by identifying anomalies in tax filings, providing early warnings to tax professionals before submission.

Beyond compliance, AI improves tax advisory services through predictive analytics and scenario modeling, enabling firms to assess tax implications and optimize tax-saving strategies. It supports audit and dispute resolution by analyzing financial data, identifying discrepancies, and preparing documentation. AI can also simulate potential dispute outcomes based on legal precedents and tax regulations (Brown et al., 2020).

Automation reduces the time and effort required for tax research, preparation, and compliance, allowing tax professionals to focus on higher-value advisory services. AI enhances accuracy by cross-checking tax data against regulations, mitigating errors and penalties. Economically, AI-driven automation lowers operational costs for accounting firms, making tax services more accessible (Kokina & Davenport, 2017). Additionally, AI-powered fraud detection tools help prevent financial losses, while real-time tax planning capabilities support informed financial decision-making.

For taxpayers, AI simplifies tax filing and planning by analyzing financial circumstances and offering guidance on deductions and benefits. It helps users navigate

complex tax regulations and provides tailored advice to optimize tax outcomes. AI's ability to identify non-compliance risks and discrepancies is invaluable to tax professionals, and its natural language processing (NLP) functions deliver real-time guidance with comprehensive explanations, empowering users to make well-informed decisions while mitigating legal risks (Munoko et al., 2020).

9.3.3.2.1 Risks in tax

Despite these advantages, AI adoption in tax practices introduces several challenges, particularly concerning regulatory compliance, data privacy, security risks, and the potential for misinterpretation of tax laws. One primary concern is data privacy and security, as AI systems require access to sensitive financial information. Firms must implement robust cybersecurity measures to protect client data from breaches and unauthorized access. Another challenge is AI model transparency and explainability. AI operates as a "black box," making it difficult for tax professionals to understand how conclusions are reached. This lack of transparency can pose obstacles in tax audit defenses and erode client trust.

Additionally, over-reliance on AI may lead to skill gaps among tax professionals. While AI enhances efficiency, it should not replace critical thinking and analytical expertise. Firms must balance automation with professional judgment and ensure continuous training for tax professionals. Moreover, AI-generated tax filings, if incorrect, can lead to legal liabilities. AI models trained on outdated or incomplete data may produce inaccurate tax recommendations, exposing firms to compliance risks.

To ensure responsible AI adoption in tax practices, accounting firms must implement strategic measures. Developing AI governance frameworks is essential to align AI applications with tax laws, ethical considerations, and regulatory requirements. Enhancing data security measures, such as encryption protocols, multi-factor authentication, and access controls, is crucial for safeguarding sensitive tax data against cyber threats. Furthermore, increasing AI transparency and explainability is necessary for auditability and validation of AI-generated tax reports by human experts.

Continuous training and skill development are also fundamental to responsible AI adoption. Firms should invest in programs that equip tax professionals with AI literacy, enabling them to leverage AI tools effectively while maintaining analytical expertise. Moreover, maintaining human oversight and ethical AI use is vital. AI should complement rather than replace human judgment in tax practices. Firms must implement checks and balances to ensure AI recommendations align with ethical standards and legal requirements.

AI offers significant advantages in tax research, compliance, advisory services, fraud detection, and efficiency. However, firms must navigate associated risks by implementing strong governance frameworks, ensuring data security, enhancing AI transparency, providing continuous professional training, and maintaining human oversight. By adopting these strategies, accounting firms can leverage AI responsibly

to optimize tax functions while ensuring compliance and ethical integrity in the evolving tax landscape.

9.4 AI in trading of financial instruments

In recent years, one area that has gained significant attention in approaches for financial investing is called quantitative or algorithmic trading. This approach gained traction in 1982 with Renaissance Technologies' quant fund, founded by Jim Simons, which leveraged mathematical models to predict price fluctuations using black-box algorithmic trading. It relied solely on quantitative analysis to pick trades.

In recent years, AI has fundamentally transformed this approach by enhancing the analysis and interpretation of vast datasets. AI-powered algorithms can sift through enormous volumes of market data, identify patterns and correlations, forecast price movements, and leverage interdependencies among various market factors, such as historical price trends and trading volumes. Through improved efficiency, accuracy, and risk management, AI-driven trading systems enable rapid data processing and execution, providing traders with a competitive edge in dynamic markets with minimal human intervention.

9.4.1 Key benefits and applications of AI in algorithmic trading

Increased efficiency
AI-powered trading systems facilitate rapid execution of trades, reducing latency and improving order execution integrity. Automation eliminates human inefficiencies, ensuring that trading decisions are implemented swiftly and accurately.

Enhanced accuracy
AI algorithms analyze large datasets with exceptional precision, allowing traders to make well-informed, data-driven decisions. The ability to process historical and real-time data enhances forecasting and market prediction accuracy.

Improved risk management
AI-driven tools effectively identify and address portfolio risks. AI algorithms assess market volatility, liquidity risks, and counterparty risks, allowing traders to manage their portfolios intelligently and protect investments from adverse market conditions.

Adaptability to market changes

AI-driven trading systems adapt to shifting market conditions, ensuring that strategies remain effective over time. By continuously learning from new data, AI algorithms refine trading approaches and enhance performance.

Advanced data and sentiment analysis capabilities

AI enables the analysis of vast and varied amounts of data, including market trends, financial news, and historical data. This capability empowers traders to identify patterns and correlations that may be overlooked by human analysts, make autonomous, data-driven decisions, adapt trading strategies dynamically to evolving market conditions, detect market anomalies such as unusual price movements or trading volumes, and maintain discipline by adhering to predefined trading rules. In addition, AI algorithms can analyze unstructured data sources such as news articles, social media posts, and market commentaries to determine market sentiment. Understanding market sentiment allows traders to anticipate market movements and adjust their trading strategies accordingly. Natural Language Processing (NLP) powered AI systems process financial news, corporate earnings reports, and analysts' notes to extract sentiment and actionable insights. The ability to rapidly analyze textual data enhances traders' responsiveness to significant market events.

Predictive modeling with pattern recognition

Predictive modeling is a critical component of AI-driven algorithmic trading. By leveraging historical and real-time data, AI models identify recurring, intricate patterns to forecast market trends. Traders can proactively adjust their strategies based on predictive insights, ensuring they remain well-positioned in volatile markets. The continuous deployment of machine learning principles enhances forecasting accuracy over time, leading to more precise market predictions.

Enhanced risk management and portfolio optimization

AI is instrumental in refining risk management practices and optimizing portfolio compositions. Advanced AI algorithms allow traders to conduct comprehensive risk assessments, including market volatility, liquidity, and counterparty risks. By integrating AI into automated trading platforms, traders can proactively identify and mitigate risks. AI-driven strategies minimize potential losses while maximizing returns through data-driven risk analysis.

9.4.2 Challenges in AI algorithmic trading

Data quality and bias

The effectiveness of AI-driven trading systems largely depends on the quality of the data they utilize. Inaccurate, incomplete, or biased data can lead to erroneous analyses, resulting in misleading trading decisions. Bias in financial datasets may arise from historical market imbalances, data selection processes, or systemic inequalities, ultimately affecting the predictive accuracy and fairness of AI models (Bengio et al., 2020). Addressing these challenges requires rigorous data preprocessing techniques, continuous validation, and bias detection mechanisms to ensure robust algorithmic decision-making.

Model interpretability

The interpretability of AI models in algorithmic trading remains a significant challenge, particularly for complex machine learning techniques such as deep learning and reinforcement learning. Many of these models function as "black boxes," making it difficult for traders and financial analysts to comprehend the rationale behind specific trading decisions. A lack of interpretability can hinder trust in AI-driven trading strategies and complicate performance evaluation and risk assessment (Doshi-Velez & Kim, 2017). Researchers advocate for explainable AI methods to enhance transparency, improving user confidence and regulatory compliance.

Regulatory compliance

AI-based trading systems operate within a highly regulated financial environment that necessitates adherence to stringent legal and ethical standards. Regulatory bodies such as the U.S. Securities and Exchange Commission (SEC) and the European Securities and Markets Authority (ESMA) emphasize transparency, fairness, and accountability in algorithmic trading. Compliance with these regulations requires firms to implement robust monitoring mechanisms, ensure algorithmic auditability, and maintain proper documentation of trading decisions (Feng et al., 2019). As AI technology evolves, regulatory frameworks must adapt to address emerging risks and maintain market integrity.

AI is reshaping the landscape of trading, equipping traders with the tools to develop sophisticated trading strategies, execute trades with remarkable precision, and manage risks adeptly. By leveraging AI-enhanced algorithms within trading platforms, market participants can achieve a competitive advantage in the fast-paced financial environment. However, to unlock the full potential of AI, it is essential to tackle existing challenges. As advancements in AI technology continue to unfold, their influence on algorithmic trading is poised to expand, heralding a new era of innovation and operational efficiency in trading strategies and practices.

9.5 AI PCs and cloud-based AI: A comparative analysis

As artificial intelligence (AI) continues to evolve, computing paradigms have adapted to optimize performance, security, and scalability. Two prominent approaches to AI computation are AI PCs and cloud-based AI. While AI PCs operate locally, cloud-based AI relies on remote servers for processing. While cloud-based AI solutions currently have the market dominance, the advancement in semiconductors is expected to make AI PC more feasible for localized AI implementations. AI PCs might be of particular importance accounting practitioners to address the data security concerns that come with the cloud-based AI. Both platforms offer unique advantages and challenges, making them suitable for different use cases. This section explores the defining characteristics of both models and their implications for AI applications.

9.5.1 AI PCs: Localized processing and enhanced security

AI PCs are designed to execute AI workloads directly on local hardware, reducing dependency on internet connectivity and external computing resources. One of the primary advantages of AI PCs is their ability to function offline. By minimizing reliance on cloud infrastructure, AI PCs ensure uninterrupted processing, making them particularly useful in scenarios where internet access is limited or unreliable.

The performance of AI PCs is fundamentally constrained by hardware specifications, including the central processing unit (CPU), graphics processing unit (GPU), and random access memory (RAM). While advancements in hardware technology have enabled AI PCs to handle increasingly complex AI models, their computational power remains limited to local system capabilities.

A notable benefit of AI PCs is their enhanced data security. Since data processing occurs locally, users retain full control over sensitive information, mitigating risks associated with cyber threats such as unauthorized access and data breaches. This feature is particularly advantageous in industries that handle confidential data, such as healthcare, finance, and government operations.

Additionally, AI PCs offer low-latency processing, as computations occur in real-time without the delays introduced by transmitting data to external servers. This is particularly beneficial for applications requiring immediate responsiveness, such as real-time analytics, financial transactions, and autonomous systems.

However, maintaining AI PCs requires manual intervention. Users or IT departments must manage hardware upgrades, software updates, and security patches, which can be resource-intensive. Furthermore, AI PCs are limited by storage capacity and scalability constraints. Large-scale AI tasks may be hindered by the physical limi-

tations of on-device storage and computing power, necessitating careful resource management.

9.5.2 Cloud-based AI: Centralized and scalable AI solutions

In contrast to AI PCs, cloud-based AI leverages remote servers hosted by providers such as Amazon Web Services (AWS), Google Cloud, and Microsoft Azure. These platforms offer virtually unlimited computational resources, enabling the execution of sophisticated AI models, including deep learning algorithms and extensive data analytics.

One of the defining features of cloud-based AI is its scalability and performance. Unlike AI PCs, which are restricted by local hardware, cloud platforms provide dynamic resource allocation, ensuring optimal performance for complex AI tasks. This scalability allows organizations to process vast datasets efficiently, making cloud AI ideal for research institutions, enterprises, and AI-driven businesses.

Cloud-based AI also supports real-time collaboration, allowing multiple users to access AI tools from any location with internet connectivity. This fosters teamwork, streamlines workflows, and facilitates seamless data sharing across geographically dispersed teams.

A key advantage of cloud-based AI is its continuous updates and maintenance. Cloud providers manage software upgrades, infrastructure enhancements, and security patches, ensuring that users benefit from the latest AI advancements without requiring manual intervention. This reduces operational burdens for organizations and enhances the longevity of AI applications.

However, cloud-based AI relies on stable internet connectivity for data processing, which can introduce latency in real-time applications. In regions with limited or unstable internet access, this dependence may pose challenges, affecting the efficiency of AI-driven workflows.

Security and compliance considerations are critical in cloud-based AI. While cloud providers implement stringent security measures such as encryption, authentication protocols, and access controls, transmitting data over the internet inherently increases exposure to cybersecurity risks. Organizations must evaluate regulatory requirements and data protection policies when adopting cloud AI solutions.

Finally, cloud-based AI follows a subscription-based cost model, typically operating on a pay-as-you-go pricing structure. While this model provides financial flexibility, expenses can accumulate over time, particularly for enterprises with high-volume data processing needs. Budget considerations and long-term cost analysis are essential for organizations leveraging cloud AI.

Both AI PCs and cloud-based AI present unique advantages and limitations, influencing their suitability for different AI applications. The choice between these platforms depends on specific use cases, budget constraints, security requirements, and

Table 9.1: Comparative Analysis: AI PC vs. Cloud-Based AI.

Feature	AI PCs	Cloud-Based AI
Processing Location	Local Machine	Remote Cloud Servers
Internet Requirement	Not Required	Required
Scalability	Limited to Hardware	Virtually Unlimited
Security & Privacy	High (Local Control)	Moderate (Cloud Dependent)
Performance Latency	Low	Potential Delay (Internet Dependency)
Maintenance	Manual	Managed by Provider
Storage Capacity	Limited	Unlimited
Collaboration	Limited	High
Cost Structure	One-Time Hardware Investment	Recurring Subscription

collaboration needs. The features of the AIPCs and Cloud-Based AI systems are compared in Table 9.1. AI PCs excel in offline functionality, data security, and low-latency processing, making them well-suited for sensitive or real-time operations. In contrast, cloud-based AI offers unparalleled scalability, collaborative capabilities, and continuous updates, catering to large-scale data analytics and enterprise-level AI deployments. Understanding these distinctions enables organizations and researchers to make informed decisions when selecting AI infrastructure for their specific requirements.

9.5.3 From AI in accounting to AI accounting agents

As discussed earlier, AI has significantly transformed the field of accounting by improving efficiency, accuracy, and strategic decision-making. Modern applications of AI leverage specialized AI agents, which are designed to perform specific accounting tasks autonomously or in collaboration with human accountants. These agents assist by analyzing vast amounts of financial data, identifying trends, and generating actionable insights that enhance decision-making.

AI agents differ from general AI in that they are task-oriented and operate based on predefined rules or learned experiences. They are designed to handle specific workflows such as invoice processing, bank reconciliation, and tax calculations with minimal human intervention. These agents can work independently or integrate with existing accounting systems, ensuring smooth automation of repetitive tasks while allowing accountants to focus on higher-level analysis and strategy.

One of the defining characteristics of AI agents is their ability to operate dynamically, responding to real-time inputs from users, financial systems, or even other AI-driven tools. Unlike traditional accounting software, AI agents are designed not just to process financial data but to take proactive actions, such as monitoring transactions, flagging anomalies, and optimizing financial workflows. This adaptability makes

them invaluable in streamlining accounting operations, reducing errors, and increasing overall productivity.

By automating routine processes and enhancing data-driven decision-making, AI agents represent a more specialized and autonomous implementation of AI. Their ability to interact dynamically within financial ecosystems ensures they can effectively support accountants, improve compliance, and enhance the accuracy of financial reporting, ultimately leading to more efficient and informed financial management.

The various accounting functions possible use through general AI versus AI Agents are explored at Table 9.2.

Table 9.2: Comparative Analysis of Use in Various Accounting Functions: AI vs. AI Agents.

Accounting Function	Role of AI	Role of AI Agent
Bookkeeping	Extracts data, categorizes transactions, and generates reports	Automates data entry and bank reconciliation
Fraud Detection	Uses anomaly detection and predictive models to flag risks	Identifies suspicious transactions based on predefined criteria
Auditing	Identifies discrepancies and irregularities through AI-powered analytics	Automates audit trail documentation and compliance checks
Financial Forecasting	Uses predictive analytics for budgeting and forecasting	Assists in generating reports and presenting financial insights
Tax Compliance	Analyzes tax laws and ensures accurate compliance	Files tax returns and automates tax calculations

9.5.4 Future of AI and AI agents in accounting

The role of AI in accounting is expected to grow significantly with advancements in machine learning, automation, and big data analytics. AI agents will become more sophisticated, capable of handling complex financial scenarios with greater autonomy. As these technologies evolve, businesses will need to balance AI's decision-making power with human oversight to ensure compliance, accuracy, and strategic financial planning.

While AI provides a broad framework for enhancing financial analysis, auditing, and decision-making in accounting, AI agents will focus on automating specific, repetitive tasks. Both technologies will play a critical role in modern accounting, enabling

firms to operate more efficiently, reduce errors, and improve financial insights. As AI continues to evolve, accountants will transition into more strategic roles, leveraging AI-driven insights for better financial management.

9.6 Conclusion

Artificial Intelligence (AI) has emerged as a transformative force in accounting, finance, and broader business operations. By automating routine processes, enhancing data analysis, strengthening fraud detection, and optimizing resource allocation, AI has significantly improved efficiency, accuracy, and strategic decision-making. The integration of AI in accounting firms, particularly in audit and tax practices, has revolutionized traditional workflows, allowing professionals to focus on higher-value advisory roles. Additionally, AI-driven financial trading and predictive analytics continue to reshape investment strategies and risk management.

Despite its numerous advantages, the adoption of AI also presents challenges, including concerns over job displacement, cybersecurity threats, regulatory compliance, and ethical considerations. The increasing reliance on AI-powered systems necessitates a balanced approach that integrates robust governance frameworks, data security measures, and human oversight. Organizations must proactively address these challenges by implementing AI risk management strategies, maintaining transparency in AI models, and ensuring ongoing professional training for finance and accounting professionals.

Looking ahead, the evolution of AI in accounting will continue to redefine industry standards, with AI accounting agents playing an increasingly autonomous role in financial management. Businesses and professionals who embrace AI's potential while mitigating its risks will be well-positioned to navigate the complexities of the digital era. As AI advances, maintaining ethical AI practices and leveraging human expertise alongside AI capabilities will be critical to sustaining long-term success and innovation in the accounting profession.

References

Bengio, Y., Lodi, A., & Prouvost, A. (2020). Machine learning for combinatorial optimization: A methodological tour d'horizon. *European Journal of Operational Research*, *290*(3), 405–421. https://doi.org/10.1016/j.ejor.2019.07.063

Brown, N., Elliott, W., & Crowley, R. (2020). What are you saying? Using topic to detect financial misreporting. Journal of Accounting Research, 58(1), 237–291.

Davenport, T. H., Brynjolfsson, E., McAfee, A., & Wilson, H. J. (2019). *Artificial intelligence: The insights you need from Harvard Business Review*. Harvard Business Press.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. https://arxiv.org/abs/1702.08608

Feng, G., Giglio, S., & Xiu, D. (2019). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *74*(5), 2207–2242. https://doi.org/10.1111/jofi.12883

Kokina, J., Blanchette, S., Davenport, T. H., & Pachamanova, D. (2025). Challenges and opportunities for artificial intelligence in auditing: Evidence from the field. *International Journal of Accounting Information Systems*, *56*, https://doi.org/10.1016/j.accinf.2025.100734

Kokina, J., & Davenport, T. H. (2017). The emergence of artificial intelligence: How automation is changing auditing. *Journal of emerging technologies in accounting*, *14*(1), 115–122.

Lakhani, K. (2023). AI won't replace humans — but humans with AI will replace humans without AI. *Harvard Business Review*, *23 (4)*. Retrieved from https://d3.harvard.edu/ai-wont-replace-humans-but-humans-with-ai-will-replace-humans-without-ai/

Lawrence, P. (2023). 3 trends that will reshape accounting and finance in 2023. Available at https://www.journalofaccountancy.com/news/2023/mar/3-trends-reshape-accounting-finance-2023.html

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. AI Magazine, 27(4), 12. https://doi.org/10.1609/aimag.v27i4.1904

Munoko, I., Brown-Liburd, H. L.,&Vasarhelyi, M. (2020). The ethical implications of using artificial intelligence in auditing. *Journal of Business Ethics*, *167*, 209–234.

The White House. (2023, October 30). *Fact sheet: President Biden issues executive order on safe, secure, and trustworthy artificial intelligence*. The White House. https://www.dhs.gov/archive/news/2023/10/30/fact-sheet-biden-harris-administration-executive-order-directs-dhs-lead-responsible

World Economic Forum. 2023. Future of Jobs Report. https://www.weforum.org/publications/the-future-of-jobs-report-2023/.

Zhao. J., Wang X. (2024) Unleashing efficiency and insights: Exploring the potential applications and challenges of ChatGPT in accounting. *Journal of Corporate Accounting and Finance*. 2024; 35, 269–276.

Ore A. Soluade and Robert J. Richardson

10 Autonomous vehicles: Challenges and concerns

10.1 Introduction

Autonomous vehicles (AVs) have the potential to significantly impact society by providing driverless mobility to a broad range of people. Those who are currently able to drive a vehicle, and those who don't due to physical, psychological or other conditions will be affected by this new technology. AVs can improve access to mobility to lower- and middle-income classes, as the costs of transportation is reduced. Some experts estimate that this form of transportation will become almost 50% cheaper compared to current costs. Additional benefits of the new technology include reduced travel time due to less traffic jams and passengers can be productive during their journey in a safer vehicle.

As with all technologies, there are pros and cons. While millions of jobs might be transformed, many will be lost as demand for human drivers in the transportation sector disappear. Some jobs such as taxi drivers will disappear completely, while other jobs such as police officers will change as they will not be giving speeding and parking tickets.

Autonomous vehicles also have secondary effects. As AVs are expected to be safer than vehicles with human drivers, the chance of getting in an accident is reduced. The reduction in accidents impacts the insurance industry and car repair centers. Car manufacturers will offer insurance services to customers, since they can offer this at a lower rate than current insurance providers. Public transportation becomes redundant, as AVs become cheaper than public transport, safer in terms of less accidents, and more comfortable in terms of privacy and hygiene.

Vehicle manufacturers are developing this technology, while the public slowly embraces the concept of autonomous driving. Stakeholders including policymakers and mobility-related industries need to understand how AVs impact society and their business. This chapter explores how autonomous vehicles have the potential to create and destroy value for society, the economy and the environment of the world.

10.2 Classification levels for autonomous vehicles

The Society of Automotive Engineers (SAE) International and automotive experts from around the world developed a classification system for defining driving automation for motor vehicles. This system was adopted by the U.S. Department of Transportation and the United Nations.

https://doi.org/10.1515/9783111595306-010

Department of Transportation's National Highway Traffic Safety Administration (NHTSA) and SAE each divide automated functionality into distinct levels. NHTSA's work focused on vehicle capabilities, while SAE categorized the technology by level of driver intervention and attentiveness required for operation (NHTSA, 2022). Their agreement was adopted using the SAE Level definitions to avoid redundancy and confusion. The six levels start with Levels 0 that has no automation to Level 5 with full automation, where the full-time performance of the Automated Driving System handles all aspects of the driving task (steering, braking, accelerating, monitoring the vehicle and roadway, as well as determining when to change lanes, turn or use signals).

10.3 Benefits of automated vehicles

Autonomous vehicles hold the potential to revolutionize transportation and bring significant benefits to society. A list of the major contributions includes:

- **Improved Safety:** AVs have the potential to significantly reduce traffic accidents caused by human error, which is a leading cause of fatalities worldwide.
- **Increased Efficiency:** AVs optimize traffic flow, reduce congestion, and minimize travel time, leading to increased productivity and reduced fuel consumption.
- **Enhanced Accessibility:** AVs provide transportation options for people with disabilities, the elderly, and those who can't drive, promoting greater independence and social inclusion. AVs provide door-to-door for passengers and first-to-last mile service for package deliveries.
- **Environmental Benefits:** AVs contribute to reduced greenhouse gas emissions through optimized routes, reduced idling time in traffic jams, and increased fuel efficiency. AVs promote better land use through reduced parking spaces and less highway lanes.
- **Economic Growth:** The development and deployment of AV technology can create new jobs and industries, stimulating economic growth.

It is important to note that the successful implementation of AVs requires careful consideration of the impact on other issues.

10.4 Acceptance of vehicles using autonomous technology

Public sentiment regarding the use of this technologies in different modes of transportation is divided. Figure 10.1 indicates that 43% say they oppose the use of driverless technology in public buses, while 34% are in favor. People are in favor of driverless

delivery vehicles (42%) or taxis and ride-sharing vehicles (41%), but about one-third are opposed.

Across each of the types of vehicles measured in this survey, there is a 20%+ level of uncertainty about how they feel using the technology. For example, 25% of Americans are unsure if the technology used in driverless passenger vehicles should be used in taxis and ride-sharing vehicles.

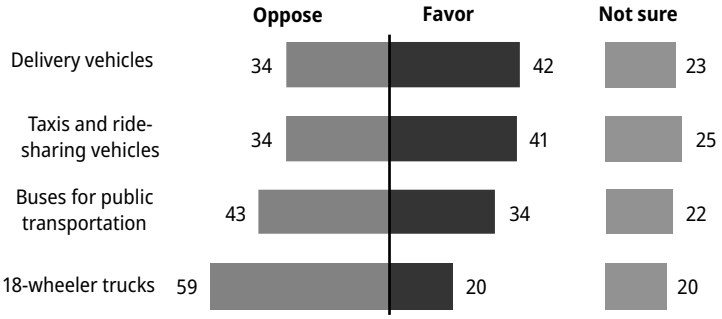


Figure 10.1: % of U.S. Adults who say they "Oppose/Favor/Not Sure" about the use of technology to operate driverless passenger vehicles (Rainie et al., 2022).

10.5 Technologies behind autonomous vehicles

There are several critical technologies that enable autonomous systems to function.

- **Sensors:** AVs use a variety of sensors to gather information about their surroundings, including:
  - LiDAR (Light Detection and Ranging): Emits laser beams to measure distances and create 3D maps of the environment.
  - Radar (Radio Detection and Ranging): Uses radio waves to detect objects, even in low visibility conditions.
  - Cameras: Capture visual information about the environment, including traffic signs, pedestrians, and other vehicles.
  - Ultrasonic Sensors: Detect objects in proximity, such as obstacles during parking.
- **Artificial Intelligence:** AI algorithms enable AVs to process sensory data, make decisions, and control their movement. These algorithms include:
  - Computer Vision: Analyzes visual data from cameras to recognize objects and traffic signs.
  - Machine Learning: Trains the AV's AI to learn from experience and improve its performance over time.
  - Deep Learning: A subset of machine learning that uses neural networks to process complex data.

Autonomous systems face significant challenges when operating in adverse weather conditions such as rain, snow, fog, and ice. These conditions can impair the performance of key sensors, each of which has specific limitations in bad weather (Harris,-ris, 2024)

Cameras: Rain and snow can obscure camera lenses, reducing visibility and contrast. In light rain, camera-based algorithms may fail to detect up to 20% of objects, with the miss rate increasing to 40% in heavier rain (Banks, 2022)

Lidar: While lidar performs well in clear conditions, it struggles in rain and snow. Raindrops and snowflakes scatter the laser signals, leading to false readings or missed detections. Spray from other vehicles can create misleading lidar targets (Teschier, 2022).

- **Radar**: Radar is more resilient in adverse weather but has limitations. It can detect objects through rain and snow but may struggle to classify smaller objects, such as pedestrians (Ansys blog, 2022)
- **Ultrasonic Sensors**: These sensors are affected by wet conditions leading to false alarms or missed detections (Tahir et al., 2024)

To enhance AV performance in adverse weather, researchers and companies are developing various solutions.

- **Sensor Fusion**: Combining data from multiple sensors can provide a more comprehensive understanding of the environment. For instance, integrating lidar, radar, and cameras can help compensate for the individual weaknesses of each sensor in bad weather (Harris, 2024).
- **Advanced Algorithms**: Developing algorithms that can adapt to changing weather conditions is crucial. These algorithms can adjust the vehicle's behavior based on real-time weather data, improving safety and reliability (Harris, 2024)
- **Ground-Penetrating Radar (GPR)**: MIT researchers are exploring the use of GPR to detect lane markings and road features beneath snow-covered surfaces. This technology allows AVs to navigate without relying solely on visible road markings, which may be obscured by snow (Beresford, 2020)
- **Sensor Protection**: Implementing physical measures, such as miniature windshield wipers for cameras and chemical deicers for sensors, can help maintain sensor clarity and functionality in snowy conditions (Banks, 2022)

It is important to note that the successful implementation of AVs requires careful consideration of the impact on cars, taxis, trucks, ethical, legal, economic and employment, environment and societal implications, as well as addressing potential challenges such as cybersecurity.

10.6 Self-driving cars

With user preference regarding the personal ownership of autonomous vehicles increasing, it is possible that the current mobility provider trend will continue as it rises in popularity. Established providers such as Uber and Lyft are already significantly present within the industry, and it is likely that new entrants will enter when business opportunities arise. In fact, 49% of businesses with fleets of between 31–100 vehicles envision driverless deliveries in the future. (Hawkins, 2018*)*.

New car design will look different and reflect the fact that driving is no longer their primary function. They may not be designed with a driver and passenger seats facing forward, but instead with a row of seats facing backward, so all the passengers in the car can more easily talk with each other. Cars may start to resemble buses or subways when it comes to passenger seat design, prioritizing comfort and efficient use of space.

More affordable taxis are in the future. The Driving Report from Ohio University indicates that with autonomous taxis, the waiting time for a cab will be reduced from the average of five minutes to under 36 seconds, and the cost of a ride will approach $0.50 per mile with a driverless car.

10.7 Autonomous trucks

Trucking is the single largest industry in the United States. More than 8.9 million people are employed in trucking, with 3.5 million of them serving as drivers. These individuals are concerned about self-driving trucks taking their jobs.

Autonomous trunks improve conditions for drivers. There is a massive truck driver shortage. The deficit estimated at around 38,000 drivers at the end of 2014. That number is expected to rise sharply, primarily because current drivers are retiring. It is predicted that industry will need to hire 89,000 drivers per year for the next decade. Autonomous trucks reduce the need for drivers since these trucks have the potential to drive for 24 hours straight stopping only to refuel. They do not need sleep breaks or lunch. This increases the efficiency of the trip and could be a big factor in reducing shipping costs.

One of the most important things about autonomous trucks is their ability to reduce fuel costs. These vehicles can accelerate, decelerate, and maintain speed without the inconsistencies seen with regular drivers. This can reduce fuel consumption by as much as 7 percent.

Another benefit of autonomous trucks is the safety improvements. According to the Federal Motor Carrier Safety Administration, there were more than 410,000 accidents involving large trucks. Of those 82,000, or 20 percent, resulted in at least one injury. Truckers are not always at fault, but things like driver fatigue, distraction, driving too close to the edge of the road, and loss of control after a blowout or other event could be stopped by self-piloting trucks.

The efficiency of self-driving trucks is another benefit for the drivers. When the company spends less on fuel, they'll be able to provide bigger paychecks for the people behind the wheel. Less fuel used means fewer emissions leading to a healthier environment.

10.8 Ethical issues

Ethical issues include how to program the vehicle to react in situations where a crash is inevitable. It is the ethical dilemma of whether a car on a collision course with a pedestrian maintains direction to save several people in the car, even if it results in killing just one person, or if the vehicle should change course to spare the life of a single pedestrian while killing the passengers. This decision is not resolved by a survey of the public, Figure 10.2. Most people in the case of an unavoidable accident want the computer system guiding the driverless car to prioritize the safety of the vehicle's passengers, rather than those outside of the vehicle (40% vs. 18%). Others are uncertain as how these systems should be programmed (41%).

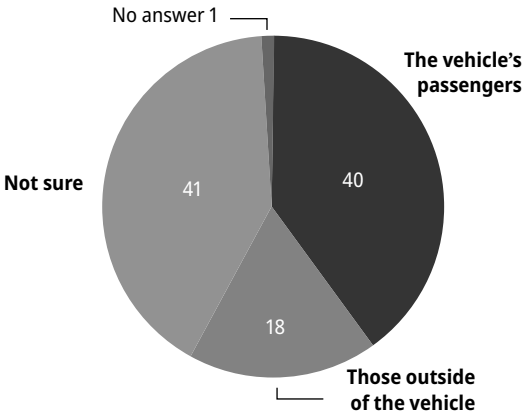


Figure 10.2: % of U.S. adults who say that in the case of an unavoidable accident, the computer systems guiding driverless passenger cars should be designed to prioritize the safety of . . . (Rainie et al., 2022).

10.9 Legal issues

Legal issues cover vehicle accidents and an individual's right to privacy. Determining liability in accidents involving autonomous vehicles can be complex, especially when the technology fails or malfunctions. These vehicles will still rely on input from the drivers, and, in some cases, the reaction won't be quick enough. In addition, big truck

and car crashes are often more serious for the other party, so fatalities could occur because of mistakes.

Technology is changing faster than the laws are written making it difficult to ascertain liability. If the driver wasn't actively piloting the truck, it wouldn't be driver mistake. Unless they were driving negligently and not paying attention when the system wanted them to take over. Most people feel that the technology shouldn't be on the road until its fully capable.

The truth is that the carmakers will likely take responsibility for any problems with their autonomous vehicles. There was a Tesla Model S that didn't recognize a semi-truck as it pulled out. The driver was watching a movie and did not react in time, causing the vehicle to crash and kill him. He and all self-driving car owners were required to sign an agreement when purchasing the vehicle to always keep his hands on the wheel. In this case, Tesla was absolved of responsibility, because it was known that this system required the driver's attention in these situations. Carmakers are currently liable in this situation, as the law stands. The risk of accidents is very low, and the technology is designed to prevent problems. Fault attached to drivers is much more complex, if they weren't in control of the vehicle.

Privacy concerns revolve around the collection of large amounts of data from the vehicle's computer. Privacy needs to be properly protected. Governments need to establish clear regulations and standards to govern the development and deployment of information. The courts will eventually rule on these issues.

10.10 Economic and employment (jobs)

Three sectors expected to see significant impact due to AV technology are workers employed in driving occupations, servicing and repair centers and insuring vehicles.

Driving occupations employ 2.9% of all workers in the US. These workers, 4.5 million, could potentially lose their jobs due to AVs. The occupations that suffer the most include delivery and heavy truck drivers, bus drivers, and taxi/chauffeur drivers. With an average income of around $45,000 per year, the total loss of income amounts to around $180 billion per year.

There could be resistance from professional drivers and unions who are threatened by job losses (Dvorak, 2015). In addition, there could be job losses in public transit services and crash repair shops. A frequently cited paper by Carl Benedikt Frey and Michael Osborne (2017) found that automated cars would make many jobs redundant. The industry has, however created thousands of jobs in low-income countries for workers who train autonomous systems (Anwar, 2020).

Insuring vehicles in the US market accounts for more than $300 billion annually (Ibisworld, 2020). The auto insurance industry exists to mitigate the risks of human drivers behind the wheel. When humans are no longer operating the vehicles, the

risk disappears and could potentially lower auto insurance prices. A shift is already taking place in the integration of insurance services. Car manufacturers are offering their own car insurance for customers, because they trust the safety of their vehicles. This is based on the current use of autonomous technologies and the expected increase of usage in the future as new autonomous technologies are being developed. A KPMG study (Albright, 2015) calculated that there is a potential for a 40% drop in total loss costs in 2040, compared to 2013. This in return will potentially reduce the auto insurance premiums benefiting users of AVs, eventually leading to a decrease in income for insurance companies and fewer employees.

Service and repair centers will see a significant change in income. In the US, vehicle crashes currently result in $71 billion a year in repair costs. Reducing the number of vehicles crashes can reduce this income by 33–90% (Albright, 2015).

Reduced revenues for municipal governments are expected, because AVs will not run red lights, speed down the highway over the limit, or overstay metered parking spots. There are no national statistic on the amount U.S. drivers pay each year in traffic violations or parking tickets. But the report points out that it adds up to hundreds of millions of dollars for some cities. This impacts the city's budget. For example, in Seattle, traffic fines constitute 2.6% of the city's operating fund. Los Angeles earned $161 million in revenue from parking tickets in 2014, while the 20 largest cities in California collect $40 million annually from vehicle towing, which is split with towing companies. States lose revenue as vehicle registrations drop. Gardner, 2015)

10.11 Environmental issues

Vehicle automation can improve fuel economy of the car by optimizing the drive cycle, as well as increasing congested traffic speeds by an estimated 8%–13% (Fagnant et al., 2015*).* Reduced traffic congestion and the improvements in traffic flow due to widespread use of automated cars will translate into higher fuel efficiency, ranging from a 23%–39% increase. Additionally, self-driving cars will be able to accelerate and brake more efficiently, meaning higher fuel economy from reducing wasted energy typically associated with inefficient changes to speed (Pyper, 2015).

With the increased demand, there may also be a reduction in the use of more sustainable modes, such as public transportation (May et al., 2020). It is expected that convenience of the automated vehicles encourages the consumers to travel more, and this induced demand may partially or fully offset the fuel efficiency improvement brought by automation. Overall, the consequences of vehicle automation on global energy demand and emissions are highly uncertain and heavily depends on the combined effect of changes in consumer behavior, policy intervention, technological progress and vehicle technology (Taiebat, 2019).

By reducing the labor and other costs of the mobility, automated cars could reduce the number of cars that are individually owned and replaced by taxis and other car-sharing services (Woodyard, 2015). This would also dramatically reduce the production of automobiles, with corresponding impact on environmental and economic effects.

A review found that private autonomous vehicles may increase total travel, whereas autonomous buses may lead to reduced car use (Mouratidis et al., 2021). The lack of stressful driving, more productive time during the trip, and the potential savings in travel time and cost could become an incentive to live far away from cities, where housing is cheaper. This increases travel distances and induces more urban sprawl, raises energy consumption and enlarges the carbon footprint of urban travel (Smith, 2015). There is the risk that traffic congestion might increase, rather than decrease. Appropriate public policies and regulations, such as zoning, pricing, and urban design are required to avoid the negative impacts of increased suburbanization and longer distance traveled (Anderson et al., 2016).

Since many autonomous vehicles are going to rely on electricity to operate, the demand for lithium batteries increases. Similarly, radar, sensors, and high-speed internet connectivity require higher auxiliary power from vehicles to generate greater power from batteries. (Taiebat et al., 2019). The larger battery requirement causes a necessary increase in the supply of these type of batteries from the chemical industry. With the expected increase of battery-powered vehicles, the petroleum industry is expected to undergo a decline in demand. As this implication depends on the adoption rate of autonomous vehicles, it is unsure to what extent this implication will disrupt this industry. This transition phase of oil to electricity allows companies to explore whether there are business opportunities for them in the new energy ecosystem. Mohan, Sripad and his team of researchers (2020) found that the electricity consumption of all the automation technology, including sensors, computation, internet access as well as the increased drag from sensors causes up to a 15% impact on the range of an automated electric vehicle, therefore, implying that the larger battery requirement might not be as large as previously assumed.

More than a quarter of all greenhouse gas emissions in the US come from the transportation sector. AV technology can improve the energy use of vehicles by improving the way vehicles brake and accelerate according to the Department of Energy resulting in a reduction of 90% in emissions (Worland, 2016).

AVs effect on the CO_2 emissions of vehicles is minimal because most of these AVs are electric. The impact of electrification on CO_2 emissions depends on the energy source from which the energy is produced: renewable versus non-renewable. Furthermore, the production of AVs will have a different environmental impact when compared to gasoline vehicles due to the use of different materials such as lithium batteries. Therefore, the full lifecycle of these vehicles should be considered to determine the actual environmental impact including greenhouse gas emissions. (Hoekstra (2019).

In Figure 10.3, the 2050 baseline energy consumption forecasts for autonomous vehicles is presented.

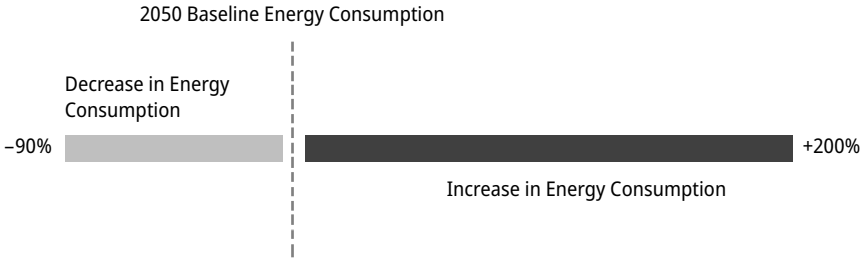


Figure 10.3: U.S. Energy Consumption in 2050 (Brown et al., 2014).

Most vehicles are also expected to be battery-electric or hybrid, resulting in an increase of electricity consumption by the transportation sector. In 2017, renewable sources made up only 11% of total energy production. Given a potential increase in electricity demand, is it to be generated using coal or natural gas resulting in increased greenhouse emissions.

10.12 Societal implications of AVs

Accident reduction was addressed by the United Nations when the topic was incorporated in a set of an ambitious goals to reduce road traffic deaths and injuries within the UN Sustainable Development Goals. According to the World Health Organization (2018), 1.35 million people died worldwide in 2016 because of traffic incidents.

In 2010, the cost of motor vehicle crashes in the United States alone was estimated at a staggering $836 billion. Of this, $242 billion was linked to economic damage, including $57.6 billion in lost workplace productivity. The cost due to loss of life and decreased quality of life due to injuries was estimated to be almost $600 billion (NHTS, 2015). Studies indicate that most of the underlying reasons for accidents are related to human errors. For example, distraction during driving is a key factor in vehicle crashes and around 30% of all road deaths in the US are alcohol related (CDC, 2020). Human behavior related accidents have the potential to be reduced significantly by autonomous technologies. Studies estimate that AVs can reduce the chance of a vehicle crash by 33% ([IIHS20]), while others indicate a 90% reduction). Despite this lack of consensus, only a 1% reduction in road accidents would result in a cost reduction of more than $8 billion annually in the US. which implies significant value to be gained from AVs for several stakeholders (McKinsey, 2015).

Congestion reduction is major issue in large cities. The global traffic scorecard of INRIX, an organization measuring traffic data, indicates that congestion and mobility trends costs citizens in the US an average 97 hours a year. The cost of this is

$87 billion in 2018, an average of $1,348 per driver. The total cost of congestion in the freight sector can be as high as $74 billon a year in the US. (McCarthy, 2020).

The more vehicles become autonomous, the more they interact. Future AVs may be capable of communicating with each other using vehicle to vehicle technology (V2V)). This enables smoother traffic flow and a higher road capacity, because vehicles can drive closer to each other resulting in less congestion. This depends on the level of market penetration by AVs. With more AVs on the road able to communicate with each other, the greater this effect will be. Eventually, platoons can be formed, which enables much closer spacing between cars (Martínez-Díaz, 2024).

Increased accessibility effects those who are not able or qualified to drive to use a fully AV. Some estimate that this will result in more vehicle miles travelled with a 14%, increase in congestion (Fagnant and Kockelman, 2015) This is because AVs could provide new mobility options for millions more Americans. There are 49 million Americans over the age of 65, and 53 million people have some type of disability. In many places, employment or independent living rests on the ability to drive. Automated vehicles extend this ability and freedom to millions more. One study (Ruderman, 2017) suggests that automated vehicles could create new employment opportunities for approximately 2 million people with disabilities.

Travel cost reduction A study by KPMG (Albright et al., 2015) revealed that the total cost per mile (including fixed and variable costs) of today's cars are $0.82. Future mobility cars could reduce that cost to $0.43 per mile, almost a 50% reduction (Albright,-bright, 2015).

Travel time not only affects congestion, but the freedom of what people can do while travelling. As fully autonomous vehicles will enable users to use their time productively during travel, this can save society significant costs. People spend 49 minutes on their commute every day. A recent study stated that automated vehicles could free up as much as 50 minutes each day per passenger in the US that had previously been dedicated to driving. It is estimated that users will only spend 75% of their travel time effectively (NHTS, 2022).

10.13 Technology: Cybersecurity and algorithms

AVs can only be autonomous if they can make decisions. The technology behind autonomous vehicles is based on computer algorithms (programs). These determine what to do when certain situations occur. For example, when a person crosses the street, the algorithm needs to first identify a moving object, then it needs to determine what it is, and thirdly it should determine if the vehicle should slow down (if possible), or steer the vehicle in another direction to avoid a crash.

Algorithms need to be programmed and trained to control an AV. The programming of basic rules is done by people; they decide what the AV will do. To test the effectiveness of these rules, the algorithm is trained by data and testing in real life situations. The results are analyzed by people who then modify the algorithm's architecture so that it makes better decisions. Autonomous vehicles may not be able to handle situations they haven't been programmed to anticipate, leading to unexpected behavior.

A recent article *Science* (Bonnefon et al., 2016) indicated that there are many challenges when defining these algorithms. Moral decisions will need to be made. Software developers, industry leaders and safety experts grapple with the modern version of this question of whose safety should be the priority in the event of an accident involving an autonomous vehicle. This is a question that some experts themselves are unsure how to answer, while others critique the usefulness of this situation altogether.

Programmers currently working on these algorithms were asked how they would deal with these situations in a survey. They indicated that they would not approve regulations mandating self-sacrifice where the car changes course killing the passengers. These regulations would make them less willing to buy an autonomous vehicle (Bonnefon et al, 2016).

Cybersecurity is an important feature in the acceptance of the AV systems. With self-driving cars controlled entirely by computers, hardware and software is vulnerable to be exploited by malicious hackers. These cyber attackers could be able to break into the vehicle's systems and take control of it. These cyberattacks could lead to dangerous situations.

Another cyber risk of driverless cars is having them all on the same network. Although it doesn't exist yet, with more self-driving cars on the road such a network would be necessary to help all the vehicles communicate with each other. If this network were hacked, it could stop all driverless cars from functioning.

10.14 Test city selection criteria

Cities are selected for robo-taxi tests based on a combination of environmental, regulatory, technological, and community factors. Major criteria are:

- **Climate and Geography**: Autonomous vehicles perform optimally in areas with mild weather and straightforward road structures. Cities like Phoenix and Austin are favored for their dry conditions and well-maintained streets, which reduce the challenges (Roth, 2023)
- **Regulatory Environment**: Supportive local and state regulations are crucial. Cities like San Francisco and Washington, D.C. have established frameworks that permit testing and, in some cases, commercial operations of driverless ve-

hicles. For instance, Waymo is preparing to launch its driverless taxi service in Washington, D.C. in 2026 (Leffer, 2023).
- **Urban Infrastructure**: The complexity of a city's road network influences AV performance. Cities with grid-like layouts and clear signage are more conducive to AV navigation (Leffer, 2023).
- **Community Engagement and Acceptance**: Public perception and acceptance are vital. Cities that engage with residents and address concerns about safety, job displacement, and technology reliability are more likely to see successful integration of robo-taxis. Protests and public opposition can arise if these issues are not adequately addressed (Robotaxi, 2025).
- **Market Potential and Strategic Partnerships**: Companies often select cities based on market size, technological infrastructure, and potential for growth. Partnerships with local entities, such as Uber's collaboration with WeRide to introduce robotaxis in 15 cities outside the U.S. and China, can facilitate smoother entry and scalability (Carter, 2025).

In summary, the selection of cities for robo-taxi deployments is a strategic decision that balances operational feasibility, regulatory support, public acceptance, and market opportunity (Mendoza, (2025).

10.15 Potential buyers of AVs

A proxy to identify those who are most likely to purchase an AV are presented in the following surveys related to (1) wanting to ride in an AV, (2) sharing the road with AVs, and (3) expressing an opinion suggesting that the widespread use of driverless passenger vehicles are good for Society

Figure 10.4 reveals a profile of U.S. adults who would want to ride in a driverless car. Most Americans, 63%, would not want to ride in a driverless passenger vehicle if they had the opportunity, while 37% say they would want to do this.

Developing a proxy for purchasing an AV based on interest in riding in a driverless vehicle varies by demographic groups. The profile includes by category: gender, a male (male 46% to female 27%), age group, people ages 18–29 with 51% and education, post graduate degrees (47%). People who have the most knowledge of AVs were most likely to have a positive impression (54%).

As with wanting to ride in a driverless vehicle, similar results were generated from Figure 10.5. The demographic profile describing the most likely person to purchase an AV is by gender, a male (male 27% to female 15%), by age group, people ages 18–29 with 31% and by education, College Graduate (27%) and post graduate degrees (25%). Those who have heard a lot about driverless cars are more likely than those who have heard nothing to think they are a good idea.

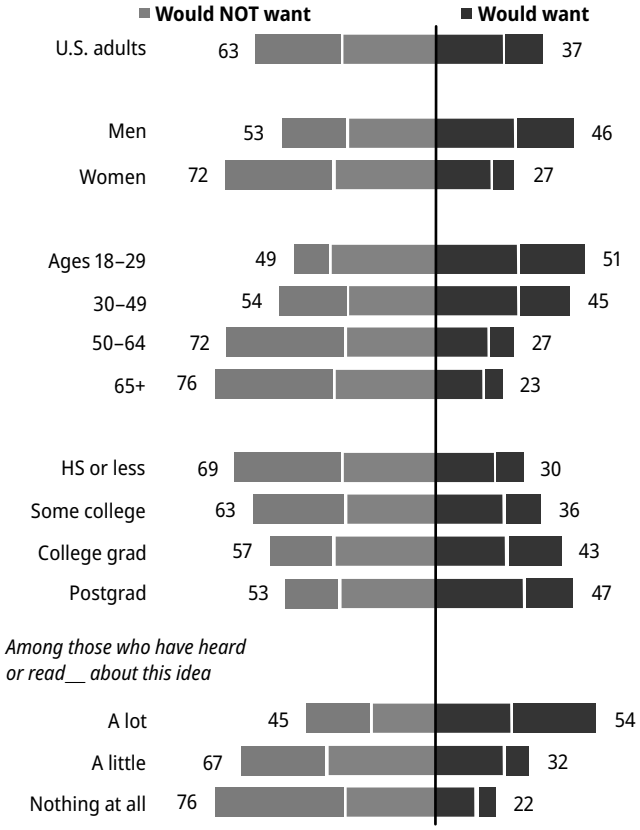


Figure 10.4: % of U.S. adults who say they definitely or probably would "Want or Not Want" to ride in a driverless passenger vehicle if they had the opportunity (Rainie et al., 2022).

In Figure 10.6, many of the groups who are more likely to say they would ride in a driverless vehicle are also more likely to say these cars are a good idea for society. For example, adults ages 18 to 49 are more likely than those 50 and older to say driverless cars are a good idea (33% vs. 19%). Men and those with more formal education are also more likely to perceive these cars as a good idea for society.

Those who have heard a lot about driverless cars (45%) are about twice as likely as those who have heard a little (22%) to say they're a good idea, and about four times as likely as those who have heard nothing at all (11%) to say the same.

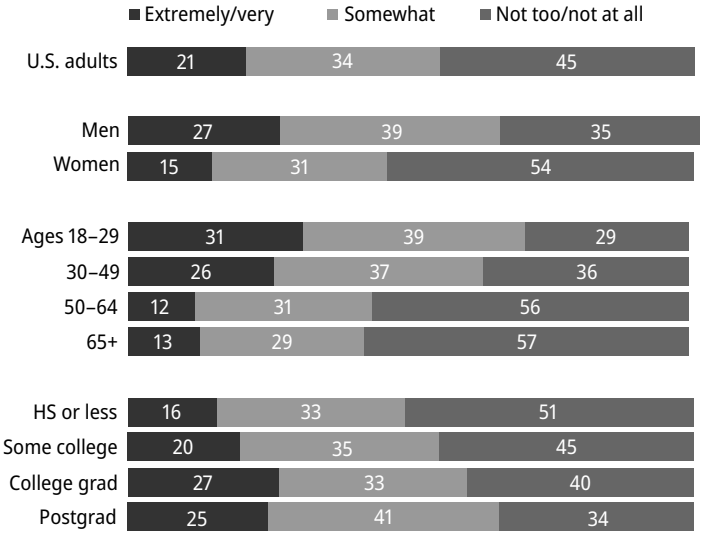


Figure 10.5: % of U.S. adults who say if the use of driverless passenger vehicles became widespread, they would feel "Extremely/Somewhat/Not too" comfortable sharing the road with AVs (Rainie et al., 2022).

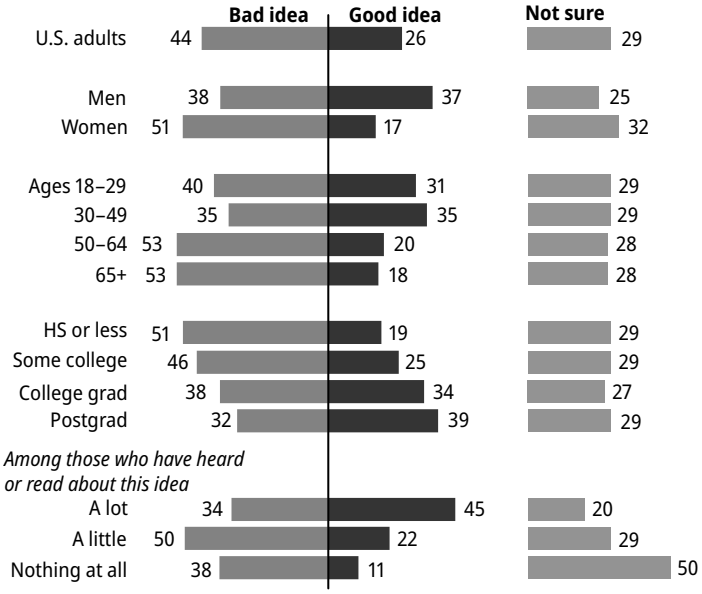


Figure 10.6: % of U.S. adults who say the widespread use of driverless passenger vehicles would be a "Good Idea/Bad Idea/Not Sure" for society (Raine et al., 2022).

10.16 Government regulation of AVs

Government role in setting driverless passenger vehicle standards is yet to be determined. However, surveys indicate different degrees to which the public would like to see the government used in setting the legislation boundaries.

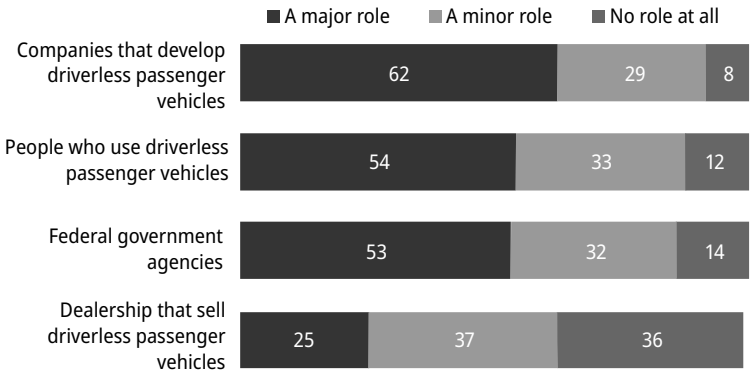


Figure 10.7: % of U.S. adults who say each of the following groups should have in setting standards for driverless passenger vehicles (Rainie et al., 2022).

To understanding the types of standards that people want to see, Americans were also asked to evaluate the level of involvement they would like to see certain groups play when it comes to setting these standards for how driverless vehicles are used in Figure 10.7.

Some 91% of Americans say the companies that develop driverless passenger vehicles should play a role in setting standards on how these vehicles are used. 62% say they should play a major role.

Similar shares say the individuals who use these driverless vehicles, and the federal government should have a major role in this process (54% and 53%, respectively). There are some differences by party when it comes to the role of the federal government in setting standards: 66% of Democrats say it should have a major role, compared with 38% of Republicans who say the same. And 20% of Republicans say the federal government should have no role at all.

Meanwhile, the public is less enthused about having car dealerships that sell driverless passenger vehicles playing a major role in setting standards with 36% saying they should have no role at all.

The U.S. Department of Transportation has been mapping safety standards and plans for introducing autonomous vehicles onto the nation's roadways. Accordingly, their survey in Figure 10.8 presents people's views about the scope of those regulations and which groups should be involved in setting the standards.

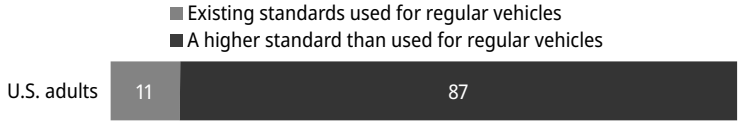


Figure 10.8: % of U.S. adults who say that when it comes to ensuring safety and effectiveness, driverless passenger vehicles should be tested using . . . (Rainie et al., 2022).

There is strong agreement among Americans that the standards used to test the safety of regular vehicles are inadequate when it comes to driverless ones. A clear majority of Americans (87%) say driverless vehicles should be tested using a higher standard than is used for regular vehicles. Only 11% believe that existing standards used for regular vehicles would be enough to ensure the safety and effectiveness of autonomous vehicles.

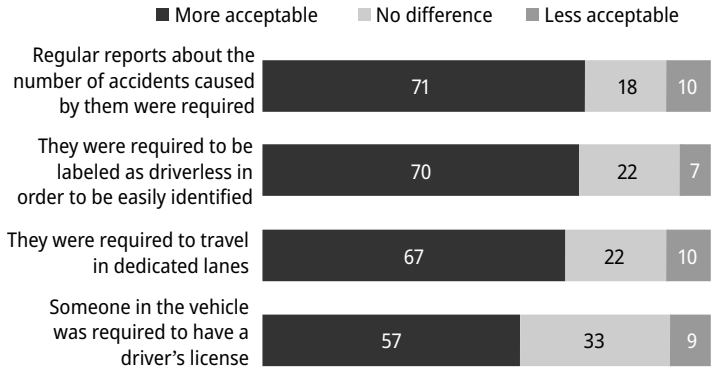


Figure 10.9: % of U.S. adults who say each of the following would make use of driverless vehicles (Rainie et al., 2022).

In Figure 10.9, Americans by 71% indicated that driverless passenger vehicles would be more acceptable if regular reports about the number of accidents caused by them were required, if autonomous cars were labeled as driverless in order to be easily identified (70%), and if such cars were required to travel in dedicated lanes (67%). Some 57% say use of driverless cars would be more acceptable if someone in the vehicle was required to have a driver's license.

10.17 Summary of challenges and concerns

Despite their potential benefits, AVs present several challenges and concerns:

- **Technical Challenges:** Developing reliable and safe AV technology requires significant advancements in sensors, AI, and software. Autonomous vehicles may struggle with adverse weather conditions or unexpected environmental factors that humans can easily handle.
- **False Sense of Security:** People may become over-reliant on autonomous systems and fail to adequately monitor the vehicle, potentially leading to dangerous situations.
- **Public Acceptance:** Public trust and acceptance are crucial for the widespread adoption of autonomous vehicles.
- **Lack of Standardization:** The lack of standardized regulations and protocols for autonomous vehicles can hinder their growth and deployment. This includes the development of a single, universal plug for all manufacturers at charging stations.
- **Infrastructure Issues**: The deployment of autonomous vehicles may require significant investments in infrastructure, including roadways, charging stations, and communication networks.
- **High Cost:** The initial cost of autonomous vehicles and the infrastructure required to support them can be substantial.
- **Impact on Driving Culture:** The shift to autonomous vehicles could diminish the enjoyment and personal freedom associated with driving.
- **Need for Human Oversight:** While autonomous vehicles aim to reduce human error, they still require human oversight, particularly in complex or unexpected situations.
- **Adaptability to Mixed Traffic:** Autonomous vehicles need to be able to navigate roads shared with human-driven vehicles, which can be challenging.

10.18 Conclusion

The future of AVs is promising, but it will require continued research, development, and collaboration among all stakeholders. As technology advances and regulations evolve, AVs are expected to gradually become a part of our transportation landscape, transforming the way we travel and live.

Autonomous vehicles have the potential to impact society significantly in the coming years. A positive factor is the number of vehicle crashes could be reduced, and travel time can be used more effectively which could result in an annual societal benefit of more than $750 billion in the US alone.

Although AVs have the potential to improve energy use of vehicles and reduce the carbon emissions, rebound effects could mitigate this effect due to an increase in use of road vehicles. This could furthermore also mitigate the positive effect AVs could have on congestion. Car sharing, flexible commuting times and working from home are therefore key considerations to optimize the impact AVs.

Many sectors will be impacted by the introduction of AVs: not only will jobs disappear in certain sectors, but there will also be a shift of jobs between different sectors. This effect may result in an annual cost of more than $350 billion annually in the US alone. Transitioning the skillsets of the affected groups is therefore key to avoid such societal costs. (Bonnefon et al., 2016)

All of these impacts are driven by the acceptance of AV technology. This acceptance will depend on people's trust in the system, including the moral decisions made when developing the algorithms used in AV technologies, as well as the (cyber)security elements and reliability of the systems. Collaboration between different disciplines and stakeholders in the autonomous vehicle space is critical in order to lead to the most optimal transitioning of this new technology.

References

Albright, J., Bali, A., Schieider, J.& Nyce, C. (2015). Marketplace of change: automobile insurance in the era of autonomous vehicles. https://assets.kpmg/content/dam/kpmg/pdf/2016/06/id-market-place-of-change-automobile-insurance-in-the-era-of-autonomous-vehicles.pdf

Anderson, J., Kalra, N., Stanley, K. Sorensen, P., Samaras, Constantine, S.& Oluwatola, T. A. (2016, May 22). Autonomous vehicle technology: guide for policymakers". *RAND Corporation*. https://www.rand.org/pubs/research_reports/RR443-2.html

Ansys Blog. (2022, February 24). How do self-driving car sensors work in snow? https://www.ansys.com/en-in/blog/self-driving-cars-sensors-snow?utm_source=chatgpt.com

Anwar, M. and Graham, M. (2020, April 20). Digital labor at economic margins: African workers and the global information economy. *Review of African Political Economy*. 47 (163): 95–105. https://ora.ox.ac.uk/objects/uuid:4f4e9109-e6cd-467a-95ca-6c4c5938fc3b

Banks, Martin (2022, January 31). Innovative AI targets the problem of drivable path detection in winter weather. https://www.roboticstomorrow.com/story/2022/01/innovative-ai-targets-the-problem-of-drivable-path-detection-in-winter-weather/18046?utm_source=chatgpt.com

Beresford, Colin. (2020, February 25). Autonomous cars struggle in snow, but MIT has a solution for that. https://www.caranddriver.com/news/a31098296/autonomous-cars-snow-mit-researchers-solution/?utm_source=chatgpt.com

Bonnefon, J. F., Shariff, A. & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, vol. 352*, *6293.* https://www.science.org/doi/10.1126/science.aaf2654

Brown, A., Jeffrey G. & Repac, B. (2014) "An Analysis of Possible Energy Impacts of Automated Vehicle," Road Vehicle Automation, Springer International Publishing (2014)

Carter, Tom (2025, May 23). Weride ceo explains how uber deal help make robotaxis safer. https://www.businessinsider.com/weride-ceo-explains-how-uber-deal-help-make-robotaxis-safer-2025-5?utm_source=chatgpt.com

CDC (2020). Impaired driving: Get the facts. https://www.cdc.gov/transportationsafety/impaired_driving/impaired-drv_factsheet.html

Dvorak, John C. (2015, September30)."There's a Bumpy Road Ahead for Driverless Cars". *PCMag*. https://www.pcmag.com/opinions/theres-a-bumpy-road-ahead-for-driverless-cars

Fagnant, D. J.; Kockelman, Kara (2015, July 1). *"Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations*". *Transportation Research Part A: Policy and Practice. 77: 167–181.*

Frey, Carl Benedikt; Osborne, Michael A. (2017, January 1). "The future of employment: How susceptible are jobs to computerization?". *Technological Forecasting and Social Change*. 114: 254–280.

Gardner, Greg (2015, February 15). Self-driving cars may lead to fewer vehicles driven more. https://www.usatoday.com/story/money/cars/2015/02/09/self-driving-cars-fewer-vehicles-driven-more/23145223/

Harris, Calvin (2024,July 19). Autonomous cars in adverse weather conditions. https://www.expertsonautos.com/autonomous-cars-weather-conditions/?utm_source=chatgpt.com

Hawkins, Andrew. (2018, January 2). You can take a ride in a self-driving Lyft during CES, the Verge. https://www.theverge.com/2018/1/2/16841090/lyft-aptiv-self-driving-car-ces-2018

Hoekstra, Auke. (2019). The Underestimated Potential of Battery Electric Vehicles to Reduce Emissions. Joule. https://www.sciencedirect.com/science/article/pii/S2542435119302715

Ibisworld (2020). Automobile insurance in the US – market size 2004-2026. Retrieved from: https://www.ibisworld.com/industry-statistics/market-size/automobile-insurance-united-states/#:~:text=The%20market%20size%2C%20measured%20by,decline%20%2D1.4%25%20in%202020

IIHS (2020). Self-driving vehicles could struggle to eliminate most crashes. https://www.iihs.org/news/detail/self-driving-vehicles-could-struggle-to-eliminate-most-crashes

Leffer, L. (2023). Robo-Taxi Takeover' Hits Speed Bumps. https://www.scientificamerican.com/article/robo-taxi-takeover-hits-speed-bumps/?utm_source=chatgpt.com

May, A; Shepherd, S; Pfaffenbichler, P & Emberger, G. (2020, November 1). "The potential impacts of automated cars on urban transport: An exploratory analysis. *Transport Policy*. 98: 127–138. m/science/article/abs/pii/S0967070X203036196/https:/ www.sciencedirect.com/science/article/abs/pii/S0967070X20303619

Martínez-Díaz, M., Haddad, C. A., Soriguera, F. & Antoniou, C. (2024, July 23). Impacts of Platooning of Connected Automated Vehicles on Highways. DOI:10.1109/TITS.2024.3350776

McCarthy, Niali (2020, March 10). Traffic congestion costs U.S. cities billions of dollars every year. https://www.forbes.com/sites/niallmccarthy/2020/03/10/traffic-congestion-costs-us-cities-billions-of-dollars-every-year-infographic/?ctpv=searchpage

McKinsey & Company (2015). Ten-way autonomous driving could redefine the automotive world. https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/ten-ways-autonomous-driving-could-redefine-the-automotive-world#

Mendoza, M. (2025, May 27). Waymo eyes Aan Antonio in Self-driving car expansion. https://www.axios.com/local/san-antonio/2025/05/27/waymo-san-antonio-cars-test-drive-downtown?utm_source=chatgpt.com

Mohan, A., Sripad, S., Vaishnav, P. & Viswanathan, V. (2020, June 20). Trade-offs between automation and light vehicle electrification. *Nature Energy* 5, 543–549 (2020). https://doi.org/10.1038/s41560-020-0644-3

Mouratidis, K., Peters, S., Van Wee, Bert *(2021, March 1).* Transportation technologies, sharing economy, and teleactivities: Implications for built environment and travel. *Transportation Research Part D: Transport and Environment*. *92*: *102716.* https://www.sciencedirect.com/science/article/pii/S1361920921000225?via%3Dihub

NHTS – National Highway Traffic Safety Administration (2015). The economic and societal impact of motor vehicle crashes, 2010. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812013

NHTSA, National Highway Traffic Safety Administration (2022). Levels of Automation. https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-03/Levels_of_Automation_Static_022822-v4-tag.pdf

Pyper, Julia (2015, September 15). Self-driving cars could cut greenhouse gas pollution". *Scientific American*.
Rainie, L., Funk, C., Anderson, M., & Tyson, A. (2022, March 17). Americans cautious about the deployment of driverless cars, Pew Research Center Report. https://www.pewresearch.org/internet/2022/03/17/americans-cautious-about-the-deployment-of-driverless-cars/
Robotaxi. (2025, May 20). In *Wikipedia*. https://en.wikipedia.org/wiki/Robotaxi?utm_source=chatgpt.com
Roth, Z. (2023, December 7). Robotaxis are coming — are cities ready? https://blog.451alliance.com/robotaxis-are-coming-are-cities-ready/?utm_source=chatgpt.com
Ruderman Family Foundation (2017). Self-driving cars: the impact on people with disabilities. Retrieved from: https://rudermanfoundation.org/white_papers/self-driving-cars-the-impact-on-people-with-disabilities/
Smith, A., & Anderson, M. (2017, October 4). Americans' attitudes toward driverless vehicles. Pew Research Center Report. https://www.pewresearch.org/internet/2017/10/04/americans-attitudes-toward-driverless-vehicles/
Smith, Noah (5 November 2015). The downside of driverless cars". *The Sydney Morning Herald* in Wikipedia. https://www.smh.com.au/business/the-downside-of-driverless-cars-20151105-gkro8y.html
Tahir, N., Zhang, Z., Asim, M., Chen, J. & Affemdo. M. (2024, February 16). Object detection in autonomous vehicles under adverse weather: a review of traditional and deep learning approaches. https://www.mdpi.com/1999-4893/17/3/103?utm_source=chatgpt.com
Taiebat, M., Brown, A., Safford, H., Qu, S. & Xu M (2019). A review on energy, environmental, and sustainability implications of connected and automated vehicles. *Environmental Science & Technology*. **52** (20): 11449–11465
Teschier, Lee. (2022, December 12). Rain dampens the autonomous vehicle parade.https://www.digikey.com/en/blog/rain-dampens-the-autonomous-vehicle-parade?utm_source=chatgpt.com
Woodyard, Chris (2015, March 5). "McKinsey study: Self-driving cars yield big benefits". *USA Today*. https://www.usatoday.com/story/money/cars/2015/03/04/mckinsey-self-driving-benefits/24382405/
Worland, J. (2016). Self-Driving Cars Could Help Save the Environment—Or Ruin It. It Depends on Us, Time. https://time.com/4476614/self-driving-cars-environment/
World Health Organization (2018). Global Status Report on Road Safety 2018. https://apps.who.int/iris/bitstream/handle/10665/277370/WHO-NMH-NVI-18.20-eng.pdf?ua=1
World Health Organization (2018). Global Status Report on Road Safety 2018. https://apps.who.int/iris/bitstream/handle/10665/277370/WHO-NMH-NVI-18.20-eng.pdf?ua=1

Vinod G. Surange and Lokpriya Gaikwad

11 Modeling AI adoption barriers for risk management in manufacturing: An ISM-MICMAC approach

11.1 Introduction

There is a significant shift occurring in the industrial sector because of the necessity for businesses to implement new technologies, particularly Artificial Intelligence (AI), to improve operational efficiency, decision-making, and Risk Management (RM) processes (Engemann & Miller, 2024). According to Peretz-Andersson et al. (2024), the ability of AI to predict future occurrences, find trends, and sift through massive amounts of data has the potential to completely revolutionize RM undertakings in the manufacturing industry. Many researchers say that using AI solutions can help managers have better control over processes because these solutions can quickly identify risks, improve the accuracy of evaluations, and offer information to monitor and lower risks. According to Baryannis et al. (2019), the industrial sector hesitates to embrace the numerous benefits of AI integration. Recent studies have highlighted challenges businesses face while integrating AI into various processes. In 2025, McKinsey & Company (Singla et al., 2025) found that compared to organizations with a smaller income, those with yearly sales of at least $500 million are seeing a faster rate of change with the integration of AI into their business functions. In a survey conducted by Gartner in 2025 (Edjlali, 2025), businesses would abandon sixty percent of their AI initiatives that are not supported by sufficient data required for AI. Technical problems, political resistance, and budgetary restrictions combine to result in these often-occurring failed efforts. These facts amply demonstrate how shallow AI penetration is in the industrial environment. Integrating AI appears challenging because of the enormous number of complicated, linked, and varied difficulties involved in RM.

Researchers Hangl et al. (2023) list operating expenses, data quality concerns, system hurdles, moral conundrums, and a lack of qualified labor as the most serious obstacles to AI adoption. Businesses have to deal with the complexity of AI models, the acceptability of stakeholders, and the prospect of combining AI with old and early-stage technologies. These complicated interdependencies incurred by integrating AI technology might make controlling industrial risks more difficult. This study used Interpretive Structural Modelling (ISM) and Cross-impact matrix multiplication applied to classification (MICMAC) analysis to examine these issues and their links. While the MICMAC analysis explores driving and dependency patterns, the ISM method reveals the hierarchical structure of these challenges. The study explores the connections between the critical barriers in AI implementation.

https://doi.org/10.1515/9783111595306-011

Furthermore, the article suggests that manufacturers, legislators, and AI field researchers will be better suited to handle these challenges and support the change toward RM driven by AI penetration. Businesses can improve their AI-integrated RM systems if they address these problematic areas. By implementing these ideas, risks will be reduced, and business continuity aspects will be improved. The findings of this study provide valuable information about the challenges preventing the widespread adoption of AI in manufacturing, as well as essential information regarding the possible way out.

11.2 Literature review

The manufacturing industry may see an uptick in efficiency and precision because of encouraging growth. Improving RM strategies by integrating AI is a promising new direction. Problems arise at every stage of RM for integrating AI, from identifying risks to monitoring and reporting on them, reducing the potential influence of AI. This limits its usefulness in RM efforts. The four main components of the literature on these challenges are identifying, evaluating, mitigating, monitoring, and reporting those risks.

11.2.1 Risk identification

11.2.1.1 Lack of interoperability standards (B1)

Every organization operates with an established set of software systems tailored to its various functional processes. Integrating newer or advanced technologies with legacy systems remains a persistent challenge due to predefined requirements and a lack of interoperability standards (Barbu et al., 2024). This hampers AI penetration and slows its implementation in RM initiatives, primarily due to the absence of standardized communication protocols. The seamless integration of AI tools into these systems is essential for a comprehensive understanding of potential risks across various operations. However, the absence of standardized communication protocols between these systems results in fragmented risk identification capabilities, limiting AI's full potential in providing a holistic view of organizational risks. According to the research findings conducted by Carvajal-Flores et al. (2024), successfully integrating and utilizing data from multiple sources depends on adopting standardized interfaces.

11.2.1.2 High Implementation Costs (B2)

The substantial upfront investment required for AI implementation is one major obstacle to their effective incorporation into risk assessments (Gurjar et al., 2024). Authors Oldemeyer et al. (2024) claim that the high initial costs of AI technologies, equipment, or experienced individuals significantly inhibit many industrial companies, especially small and medium-sized businesses (SMEs), from deploying these high-end technologies. The expenses include the new technology installation expenses and staff member training expenses on maximizing AI-powered technologies. Many companies hesitate to adopt AI for risk assessment, often sticking to traditional methods. This is despite AI's potential benefits, such as improved accuracy and more efficient RM. The high initial costs and uncertainty about AI's effectiveness in practice contribute to this reluctance.

11.2.1.3 Data Quality Issues (B3)

Companies should undertake three activities to ensure data accuracy in AI systems: data cleansing, standardizing, and continuous data governance (Whang et al., 2023). Inaccurate or inadequate data limits AI's ability to learn from prior events, affecting the accuracy of risk estimates. AI systems need data that is relevant, consistent, and of high quality if they are to operate well. AI's capacity to identify specific risk patterns might be affected by mistakes, a lack of data, or out-of-date information; thus, these factors could influence possibly erroneous decision-making (Plathottam et al., 2023).

11.2.2 Risk assessment

11.2.2.1 Insufficient Stakeholder Buy-in (B4)

The absence of endorsement from prominent figures such as senior management and department leaders is a significant barrier to using AI in risk assessment and other decision-making scenarios, according to Lada et al. (2023). This might be a significant impediment. Assessing a company's readiness to engage with AI in the context of risk assessments is essential. Stakeholders may oppose AI integration due to concerns about job displacement, lacking trust in AI, and resistance to combining traditional methods with new technologies (Cheng et al., 2023). AI's efficacy in risk analysis may likely suffer without comprehensive organizational support. Contemporary technology typically necessitates significant modifications in the business's operations and culture.

11.2.2.2 Complexity of AI Models (B5)

The risk evaluation method could be complicated for users without technical knowledge, as AI models are often perceived to be complex and challenging to understand. Authors Ångström et al. (2023) Conducted multiple investigations into the technological intricacy of AI models. Also included in this research is the challenge that non-technical stakeholders have when trying to make sense of AI-generated estimates. Decision-makers may find it challenging to comprehend the outcomes of these evaluations despite AI's ability to produce highly accurate risk measurements because algorithms, in general, and deep learning models, in particular, are unreliable. According to Mazhar et al.'s (2023) research, one approach to address this issue could be to use simpler models or mix AI with decision-support expert systems.

11.2.2.3 Ethical concerns (B6)

AI risk assessments should not be carried out without carefully examining the societal ramifications. Uniqueness and equality are key components of the discussion around these repercussions. Numerous researches demonstrate that by extending preconceptions found in prior data, AI systems may generate inaccurate assessments and inaccurate analyses of scenarios. People have trouble believing because AI decision-making evaluations provide contradictory conclusions (Ferrara, 2024; A. Sinha et al., 2024). This is especially crucial in sectors like health and safety. Organizations must first develop ethical standards and practices to address these concerns effectively and eradicate bias in developing AI systems.

11.2.2.4 Lack of transparency in AI models (B7)

It is more challenging to explain the opaque nature of AI models' prediction behavior, especially when users lack a clear understanding of risk evaluation. When there is a lack of communication between the parties, the parties are more reluctant to each other, leading to trust concerns (Sharma, 2024). It is more difficult to accept and evaluate the outcomes of AI as a consequence of this. Several studies emphasize the importance of transparent AI techniques, which aim to make AI decisions more accountable and understandable for users (Cheong, 2024; Fehr et al., 2024).

11.2.3 Risk mitigation

11.2.3.1 Lack of skilled workforce (B8)

Deploying and managing AI systems requires expertise. Researchers Surange et al. (2024) say that data science industry 4.0 technologies and AI risk minimization are complex as few competent professionals are available as experts due to competence gaps in these fields. According to researchers Deiva Ganesh & Kalpana (2022), most organizations struggle to find experienced AI model builders who can reduce key risks. Many businesses are dealing with this kind of issue. Organizations are implementing training initiatives to enhance the workforce's skills, addressing a shortage in this area. This will assist in lowering the risks connected to AI integration.

11.2.3.2 Scalability challenges (B9)

Risk-reducing technologies driven by AI have to consider scalability. Researchers Fan et al. (2023) and Zhang et al. (2021) claim that AI systems proved successful in pilot projects but could have challenges scaling their operations in significantly more prominent and complicated environments. Problems with data infrastructure, system performance, and interaction with present processes could limit the scalability of AI systems. Large-scale applications could thus not be able to use these systems.

11.2.3.3 Regulatory and compliance issues (B10)

Legal and compliance responsibilities often hinder the application of AI in RM (Mennella et al., 2024). Authors Roberts et al.'s (2024) research points to the need for global AI governance initiatives as national rules controlling data privacy, security, and technology use possibly make applying AI-powered solutions more difficult. The adoption of AI in several sectors, including finance and healthcare, is hampered by the need for companies to create extra protections and certifications to comply with these laws.

11.2.4 Risk monitoring & reporting

11.2.4.1 Data Privacy Concerns (B11)

Data privacy becomes a significant issue since AI depends on large datasets (Golda et al., 2024). Continuous, real-time risk monitoring may call for access to sensitive data, generally controlled by strict privacy standards. The utilization of personal data

and the occurrence of data breaches force businesses to avoid altogether including AI in their RM structures.

11.2.4.2 Cybersecurity Risks (B12)

The vulnerability of AI systems to hacking jeopardizes the validity of risk monitoring and reporting. Researchers Umer et al. (2024) include many different studies citing the advent of cloud computing, which has made industrial organizations vulnerable to cyberattacks. These kinds of attacks may involve adversaries modifying AI models to produce risk estimates that are either wrong or misleading. There exist loopholes in cybersecurity that make AI less reliable and trustworthy when it comes to RM. To protect AI systems from hackers, strict security standards are needed.

11.2.4.3 Lack of Real-time Capabilities (B13)

Risk monitoring systems that AI drives can interpret data in real time, speeding up the identification and response to potential dangers. According to AlMahasneh et al. (2024), the inability of many AI systems to interpret data in real-time/algorithm constraints is a significant limitation on their efficacy in quickly changing situations. The potential advantages of AI in proactive risk assessment and mitigation are undermined when data is not examined in real time, which is a must.

11.2.4.4 Integration with Legacy Systems (B14)

Integrating AI with pre-existing systems is a significant obstacle that must be overcome before it can be used in RM. Several manufacturing companies, according to S. Sinha & Lee (2024), continue to rely on age-old systems that are not compatible with the AI technologies that are currently available. Because of this incompatibility makes incorporating AI solutions into pre-existing RM systems more difficult, resulting in expensive system additions or replacements. In Table 11.1, we outline the obstacles that stand in the way of the widespread use of AI in RM specific to the manufacturing environment.

Table 11.1: Barriers to AI Adoption for RM.

Phase	Code	Obstacle	Description	Reference
Risk Identification (I)	B1	Lack of Interoperability Standards	Difficulty in integrating AI tools across systems hampers identifying risks comprehensively.	(Bag et al., 2022; Ben Ayed & Hanana, 2021; Kusiak, 2017; Lee et al., 2021; Nair et al., 2024)
	B2	High Implementation Costs	Initial costs deter organizations from investing in AI-driven risk identification.	
	B3	Data Quality Issues	Poor data quality limits AI's ability to identify accurate risk patterns.	
Risk Assessment (A)	B4	Insufficient Stakeholder Buy-in	Without organizational support, AI's potential in risk evaluation is underutilized.	(Hangl et al., 2023; Morley et al., 2023; Nair et al., 2024; von Eschenbach, 2021)
	B5	Complexity of AI Models	Complex models make assessing identified risks challenging for non-technical users.	
	B6	Ethical Concerns	Concerns over bias in AI models undermine confidence in risk assessments.	
	B7	Lack of Transparency in AI Models	Black-box models make risk evaluations difficult to justify and trust.	

(continued)

Table 11.1 (continued)

Phase	Code	Obstacle	Description	Reference
Risk Mitigation (M)	B8	Lack of Skilled Workforce	A gap in AI expertise hinders the implementation of effective risk mitigation strategies.	(Govindan, 2022; A. Sinha et al., 2024; Tariq et al., 2021)
	B9	Scalability Challenges	Difficulty scaling AI systems affects their utility in widespread risk reduction.	
	B10	Regulatory and Compliance Issues	Regulatory constraints limit deploying AI for proactive RM.	
Risk Monitoring & Reporting (R)	B11	Data Privacy Concerns	Privacy issues limit AI's access to real-time data for ongoing risk tracking.	(Bag et al., 2022; Govindan, 2022; Lee et al., 2021; Surange et al., 2021)
	B12	Cybersecurity Risks	AI systems can become targets of cyberattacks, jeopardizing monitoring.	
	B13	Lack of Real-time Capabilities	The inability to process risks in real-time weakens proactive monitoring.	
	B14	Integration with Legacy Systems	Legacy systems hinder AI's ability to function seamlessly for continuous monitoring.	

11.3 Analysis of barriers using ISM-MICMAC methodology

ISM explains the relationships between the multiple elements included within a complex network. This approach systematically arranges factors directly or indirectly connected to a system (Satvik et al., 2025; Singh Patyal et al., 2021). Implementing the ISM method starts with identifying the components connected to the problem. Developing a Structural self-interaction matrix (SSIM) depicts the interrelations among several elements after expert advice. The initial reachability matrix (IRM) is derived further and is utilized to generate an ISM model in which assertions are replaced for the nodes. Five Indian manufacturing sector decision-makers were selected for their in-depth understanding of RM and technology implementation. This ensured the topic was relevant and quality input was obtained. Each expert had over ten years of industry experience and was directly involved in strategic decision-making roles. Their selection was based on criteria such as leadership roles in operational risk, exposure to tech-driven systems, and hands-on experience with industrial automation and compliance challenges. Most manufacturing firms they represent adhere to quality management (QM) certifications such as ISO standards, where RM is a key evaluation criterion. Hence, these experts already possess a foundational understanding of RM frameworks.

Additionally, given the rapid evolution of industrial technologies, these managers are accustomed to integrating emerging tools into existing processes. Their roles inherently require a sound understanding of technological advancements and operational risk dynamics. The sample size of five is considered sufficient for the multi-criteria decision-making methodology, as supported by prior studies (Rezaei et al., 2018; Surange et al., 2024; Surange & Bokade, 2024), due to the qualitative nature of the analysis. Expert opinions were gathered through structured interviews and iterative feedback sessions to ensure consistency and depth in developing the self-interaction matrix. Table 11.2 depicts the characteristics of industry decision-makers who were consulted for their input in the process.

11.3.1 Construction of structural self-interaction matrix

An SSIM exemplifies the relationships between any two arbitrary components within the matrix. The symbols denoting these relationships are as follows: If element ‘i’ influences element ‘j’: V If element ‘j’ influences element ‘i’: A If both element ‘i’ and element ‘j’ mutually influence each other: X If there is no relationship between element ‘i’ and element ‘j’: O.

A consensus-based approach addressed expert opinion differences while constructing the SSIM. Experts provided pairwise evaluations for each barrier relationship, and the most prominent relation among their responses was selected to represent the final input. Where initial judgments diverged significantly, structured group

Table 11.2: Decision Makers Profile.

Decision Maker (DM)	Designation	Total experience in years	Manufacturing industry type
1	Mid-level Manager – Operations	12	Automotive Parts and Components
2	Senior Manager – Production	22	Electrical Equipment and Systems
3	Factory operations head	15	Clean Energy and Renewable Technology
4	Technical Lead – Production	13	Thermal Systems and Industrial Vessels
5	Executive Leader – Systems Strategy	33	Custom Metal Engineering and Fabrication

discussions were conducted in iterative rounds to deliberate and reach mutual agreement. This process ensured consistency, collective validation, and methodological rigor in accurately capturing the relationships between variables for the ISM analysis. Refer to Table 11.3, which was constructed after consulting the expert's input.

Table 11.3: Structural Self-Interaction Matrix (SSIM).

	B14	B13	B12	B11	B10	B9	B8	B7	B6	B5	B4	B3	B2	B1
B1	V	V	O	V	O	O	O	O	O	V	O	V	O	–
B2	V	V	V	O	V	O	V	V	V	V	V	O	–	
B3	V	V	O	X	O	V	O	O	O	V	O	–		
B4	V	V	X	O	V	V	V	V	V	V	–			
B5	A	A	A	A	A	A	A	A	A	–				
B6	V	V	A	O	X	V	X	X	–					
B7	V	V	A	O	X	V	X	–						
B8	V	V	A	O	X	V	–							
B9	X	X	O	A	A	–								
B10	V	V	A	O	–									
B11	V	V	O	–										
B12	O	V	–											
B13	X	–												
B14	–													

11.3.2 Formation of the initial reachability matrix

The Initial Reachability Matrix (IRM) was constructed in the following manner: A value of 1 is assigned to the reachability matrix, and the (j, i) entry is assigned a value of 0 if the (i, j) element in the SSIM matrix is equivalent to V. Whenever the (i, j) ele-

ment in the self-interaction matrix is represented by the letter A, the reachability matrix will record a value of 0 and the (j, i) entry will have a value of 1. If the element (i, j) in the self-interaction matrix is represented by X, then the entries (i, j) and (j, i) in the reachability matrix are both represented by the value 1. A value of 0 is assigned to the reachability matrix when the (i, j) element in the self-interaction matrix is represented by the letter O. Additionally, the (j, i) entry remains unchanged at 0.

To illustrate the SSIM-to-IRM conversion process using specific barrier descriptions, consider the relationship between B1 (Lack of Interoperability Standards) and B14 (Integration with Legacy Systems). Since the SSIM entry is marked as 'V,' it indicates that interoperability issues are perceived to influence the integration challenges with legacy systems. As a result, the IRM assigns a value of 1 to the cell (B1, B14) and 0 to the cell (B14, B1). In another case, For B5 (Complexity of AI Models) and B14 (Integration with Legacy Systems), the 'A' in SSIM means B14 (Integration with Legacy Systems) influences B5 (Complexity of AI Models). Legacy systems often create obstacles for newer technologies like AI, increasing the complexity of building and deploying AI models that must operate within those old infrastructures. Therefore, in the IRM, (B14, B5) is assigned a 1, while (B5, B14) is assigned a 0. For mutual relationships, such as between B3 (Data Quality Issues) and B11 (Data Privacy Concerns), the SSIM entry is marked 'X,' suggesting a bidirectional influence. For example, poor data quality can affect privacy safeguards. At the same time, privacy-related constraints may restrict data availability or structure. In this case, both (B3, B11) and (B11, B3) are assigned a value of 1 in the IRM. Conversely, if two barriers have no discernible influence on each other, such as B1 (Lack of Interoperability Standards) and B12 (Cybersecurity Risks), the SSIM entry is 'O', and thus both (B1, B12) and (B12, B1) are assigned 0 in the reachability matrix.

Refer to Table 11.4, which presents IRM.

11.3.3 Incorporating transitivity

If one relationship is shown by the phrase 'I leads to j' and another is indicated by the phrase 'j leads to k,' then by means of the transitive property one may infer 'i leads to k.' Usually, the Final Reachability Matrix shows a transitive relationship. This link shows that the impact or result of component 'i' indirectly reaches element 'k' via the intermediary element 'j' inside the hierarchical structure produced by the interactions. These transitive interactions are fundamental to understanding the numerous connections and interdependencies that exist inside a system.

To ensure structural consistency, transitivity checks were applied. For example, if barrier B2 influences B4, and B4 influences B9, then it is inferred that B2 also influences B9 indirectly. These transitive links were integrated into the matrix using iterative updates until the structure stabilized.

In order to view the final reachability matrix, please refer to Table 11.5.

Table 11.4: Reachability Matrix (RM).

Variables	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	Driving Power
B1	1	0	1	0	1	0	0	0	0	0	1	0	1	1	**6**
B2	0	1	0	1	1	1	1	1	0	1	0	1	1	1	**10**
B3	0	0	1	0	1	0	0	0	1	0	1	0	1	1	**6**
B4	0	0	0	1	1	1	1	1	1	1	0	1	1	1	**10**
B5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	**1**
B6	0	0	0	0	1	1	1	1	1	1	0	0	1	1	**8**
B7	0	0	0	0	1	1	1	1	1	1	0	0	1	1	**8**
B8	0	0	0	0	1	1	1	1	1	1	0	0	1	1	**8**
B9	0	0	0	0	1	0	0	0	1	0	0	0	1	1	**4**
B10	0	0	0	0	1	1	1	1	1	1	0	0	1	1	**8**
B11	0	0	1	0	1	0	0	0	1	0	1	0	1	1	**6**
B12	0	0	0	1	1	1	1	1	0	1	0	1	1	0	**8**
B13	0	0	0	0	1	0	0	0	1	0	0	0	1	1	**4**
B14	0	0	0	0	1	0	0	0	1	0	0	0	1	1	**4**
Dependence Power	**1**	**1**	**3**	**3**	**14**	**7**	**7**	**7**	**10**	**7**	**3**	**3**	**13**	**12**	

Table 11.5: Final Reachability Matrix (FRM).

Variables	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	Driving Power
B1	1	0	1	0	1	0	0	0	1*	0	1	0	1	1	**7**
B2	0	1	0	1	1	1	1	1	1*	1	0	1	1	1	**11**
B3	0	0	1	0	1	0	0	0	1	0	1	0	1	1	**6**
B4	0	0	0	1	1	1	1	1	1	1	0	1	1	1	**10**
B5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	**1**
B6	0	0	0	0	1	1	1	1	1	1	0	0	1	1	**8**
B7	0	0	0	0	1	1	1	1	1	1	0	0	1	1	**8**
B8	0	0	0	0	1	1	1	1	1	1	0	0	1	1	**8**
B9	0	0	0	0	1	0	0	0	1	0	0	0	1	1	**4**
B10	0	0	0	0	1	1	1	1	1	1	0	0	1	1	**8**
B11	0	0	1	0	1	0	0	0	1	0	1	0	1	1	**6**
B12	0	0	0	1	1	1	1	1	1*	1	0	1	1	1*	**10**
B13	0	0	0	0	1	0	0	0	1	0	0	0	1	1	**4**
B14	0	0	0	0	1	0	0	0	1	0	0	0	1	1	**4**
Dependence Power	**1**	**1**	**3**	**3**	**14**	**7**	**7**	**7**	**13**	**7**	**3**	**3**	**13**	**13**	

11.3.4 Partition of the reachability matrix into different levels

The final reachability matrix is analyzed to determine each factor's reachability and antecedent set (Warfield, 1974). When it comes to a particular factor, the reachability

set (R_i) is comprised of the factor itself as well as other variables that it has the potential to impact. On the other hand, the antecedent set (A_i) comprises the factor itself in addition to other elements that have the potential to influence it. By studying the common numbers in both sets, it is possible to discover an additional intersection set, which is denoted as IS = $R_i \cap A_i$. The level partitioning was carried out for the identical reachability set and the intersection set, and level 1 was allocated and placed at the top of the ISM model. Following the marking of level 1, the associated factor was left out for the net iterations, and the process was repeated as many times as necessary for all remaining factors (Refer to Table 11.6 to 11.10).

Table 11.6: Level Partitioning Iteration 1.

Elements (Mi)	Reachability Set R(Mi)	Antecedent Set A(Ni)	Intersection Set R(Mi) ∩ A(Ni)	Level
1	1, 3, 5, 9, 11, 13, 14,	1,	1,	
2	2, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14,	2,	2,	
3	3, 5, 9, 11, 13, 14,	1, 3, 11,	3, 11,	
4	4, 5, 6, 7, 8, 9, 10, 12, 13, 14,	2, 4, 12,	4, 12,	
5	5,	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,	5,	1
6	5, 6, 7, 8, 9, 10, 13, 14,	2, 4, 6, 7, 8, 10, 12,	6, 7, 8, 10,	
7	5, 6, 7, 8, 9, 10, 13, 14,	2, 4, 6, 7, 8, 10, 12,	6, 7, 8, 10,	
8	5, 6, 7, 8, 9, 10, 13, 14,	2, 4, 6, 7, 8, 10, 12,	6, 7, 8, 10,	
9	5, 9, 13, 14,	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14,	9, 13, 14,	
10	5, 6, 7, 8, 9, 10, 13, 14,	2, 4, 6, 7, 8, 10, 12,	6, 7, 8, 10,	
11	3, 5, 9, 11, 13, 14,	1, 3, 11,	3, 11,	
12	4, 5, 6, 7, 8, 9, 10, 12, 13, 14,	2, 4, 12,	4, 12,	
13	5, 9, 13, 14,	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14,	9, 13, 14,	
14	5, 9, 13, 14,	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14,	9, 13, 14,	

All variables on the same level are grouped in the conical matrix (see Table 11.11).

In order to construct a structural ISM model from a conical matrix, arrows, and nodes are considered. Both the digraph and the ISM model are seen in Figure 11.1 and 11.2.

To comprehend and analyze the driving power and dependence power of the barriers, MICMAC analysis is utilized extensively. The Driving Power and Dependence Power values of each barrier were used to generate four distinct clusters, which are referred to as autonomous, dependent, linkage, and independent, respectively (refer to Figure 11.3).

Table 11.7: Level Partitioning Iteration 2.

Elements (Mi)	Reachability Set R(Mi)	Antecedent Set A(Ni)	Intersection Set R(Mi) ∩ A(Ni)	Level
1	1, 3, 9, 11, 13, 14	1	1	
2	2, 4, 6, 7, 8, 9, 10, 12, 13, 14	2	2	
3	3, 9, 11, 13, 14	1, 3, 11	3, 11	
4	4, 6, 7, 8, 9, 10, 12, 13, 14	2, 4, 12	4, 12	
6	6, 7, 8, 9, 10, 13, 14	2, 4, 6, 7, 8, 10, 12	6, 7, 8, 10	
7	6, 7, 8, 9, 10, 13, 14	2, 4, 6, 7, 8, 10, 12	6, 7, 8, 10	
8	6, 7, 8, 9, 10, 13, 14	2, 4, 6, 7, 8, 10, 12	6, 7, 8, 10	
9	9, 13, 14	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14	9, 13, 14	2
10	6, 7, 8, 9, 10, 13, 14	2, 4, 6, 7, 8, 10, 12	6, 7, 8, 10	
11	3, 9, 11, 13, 14	1, 3, 11	3, 11	
12	4, 6, 7, 8, 9, 10, 12, 13, 14	2, 4, 12	4, 12	
13	9, 13, 14	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14	9, 13, 14	2
14	9, 13, 14	1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14	9, 13, 14	2

Table 11.8: Level Partitioning Iteration 3.

Elements (Mi)	Reachability Set R(Mi)	Antecedent Set A(Ni)	Intersection Set R(Mi) ∩ A(Ni)	Level
1	1, 3, 11	1	1	
2	2, 4, 6, 7, 8, 10, 12	2	2	
3	3, 11	1, 3, 11	3, 11	3
4	4, 6, 7, 8, 10, 12	2, 4, 12	4, 12	
6	6, 7, 8, 10	2, 4, 6, 7, 8, 10, 12	6, 7, 8, 10	3
7	6, 7, 8, 10	2, 4, 6, 7, 8, 10, 12	6, 7, 8, 10	3
8	6, 7, 8, 10	2, 4, 6, 7, 8, 10, 12	6, 7, 8, 10	3
10	6, 7, 8, 10	2, 4, 6, 7, 8, 10, 12	6, 7, 8, 10	3
11	3, 11	1, 3, 11	3, 11	3
12	4, 6, 7, 8, 10, 12	2, 4, 12	4, 12	

Table 11.9: Level Partitioning Iteration 4.

Elements(Mi)	Reachability Set R(Mi)	Antecedent Set A(Ni)	Intersection Set R(Mi) ∩ A(Ni)	Level
1	1	1	1	4
2	2, 4, 12	2	2	
4	4, 12	2, 4, 12	4, 12	4
12	4, 12	2, 4, 12	4, 12	4

Table 11.10: Level Partitioning Iteration 5.

Elements (Mi)	Reachability Set R(Mi)	Antecedent Set A(Ni)	Intersection Set R(Mi) ∩ A(Ni)	Level
2	2, 4, 12	2	2	5

Table 11.11: Conical Matrix (CM).

Variables	B5	B9	B13	B14	B3	B6	B7	B8	B10	B11	B1	B4	B12	B2	Driving Power	Level
B5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	**1**
B9	1	1	1	1	0	0	0	0	0	0	0	0	0	0	4	**2**
B13	1	1	1	1	0	0	0	0	0	0	0	0	0	0	4	**2**
B14	1	1	1	1	0	0	0	0	0	0	0	0	0	0	4	**2**
B3	1	1	1	1	1	0	0	0	0	1	0	0	0	0	6	**3**
B6	1	1	1	1	0	1	1	1	1	0	0	0	0	0	8	**3**
B7	1	1	1	1	0	1	1	1	1	0	0	0	0	0	8	**3**
B8	1	1	1	1	0	1	1	1	1	0	0	0	0	0	8	**3**
B10	1	1	1	1	0	1	1	1	1	0	0	0	0	0	8	**3**
B11	1	1	1	1	1	0	0	0	0	1	0	0	0	0	6	**3**
B1	1	1*	1	1	1	0	0	0	0	1	1	0	0	0	7	**4**
B4	1	1	1	1	0	1	1	1	1	0	0	1	1	0	10	**4**
B12	1	1*	1	1*	0	1	1	1	1	0	0	1	1	0	10	**4**
B2	1	1*	1	1	0	1	1	1	1	0	0	1	1	1	11	**5**
Dependence Power	14	13	13	13	3	7	7	7	7	3	1	3	3	1		
Level	**1**	**2**	**2**	**2**	**3**	**3**	**3**	**3**	**3**	**3**	**4**	**4**	**4**	**5**		

11.4 Results and discussion

In utilizing AI to manage production hazards, numerous obstacles exist to overcome at each stage of the RM process. An ISM method clarifies this point and provides evidence that supports it through investigations. The ISM analysis reveals that High Implementation Costs (B2) hold the highest driving power and sit at the base level (Level 5), establishing it as the most fundamental driving barrier. It significantly influences nearly all other barriers and must be addressed first for successful AI adoption in RM. At Level 4, the lack of interoperability standards (B1), insufficient stakeholder buy-in (B4), and cybersecurity risks (B12) also exhibit strong driving power and act as critical challenges. Barriers such as Ethical Concerns (B6), Lack of Transparency in AI Models (B7), Lack of Skilled Workforce (B8), and Regulatory and Compliance Issues (B10) appear at Level 3 with moderate driving and dependence powers, classifying them as linkage barriers. These are both influenced by foundational challenges and, in turn,

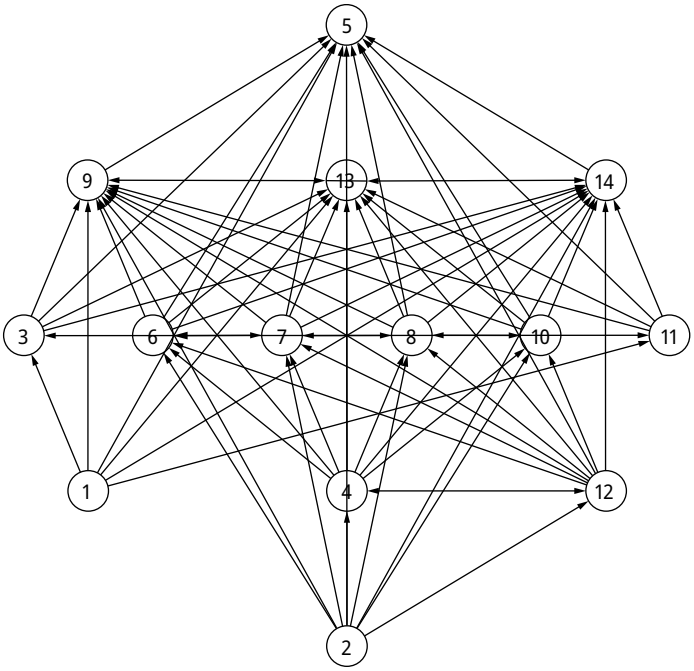


Figure 11.1: Diagraph.

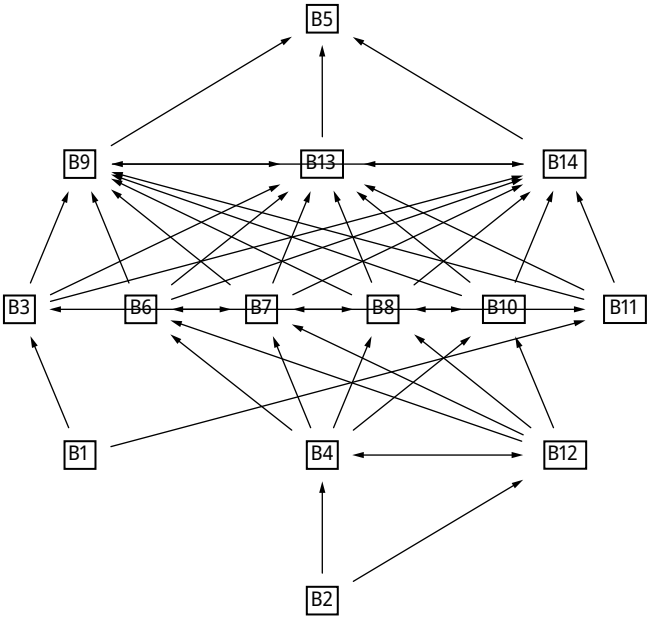


Figure 11.2: ISM model.

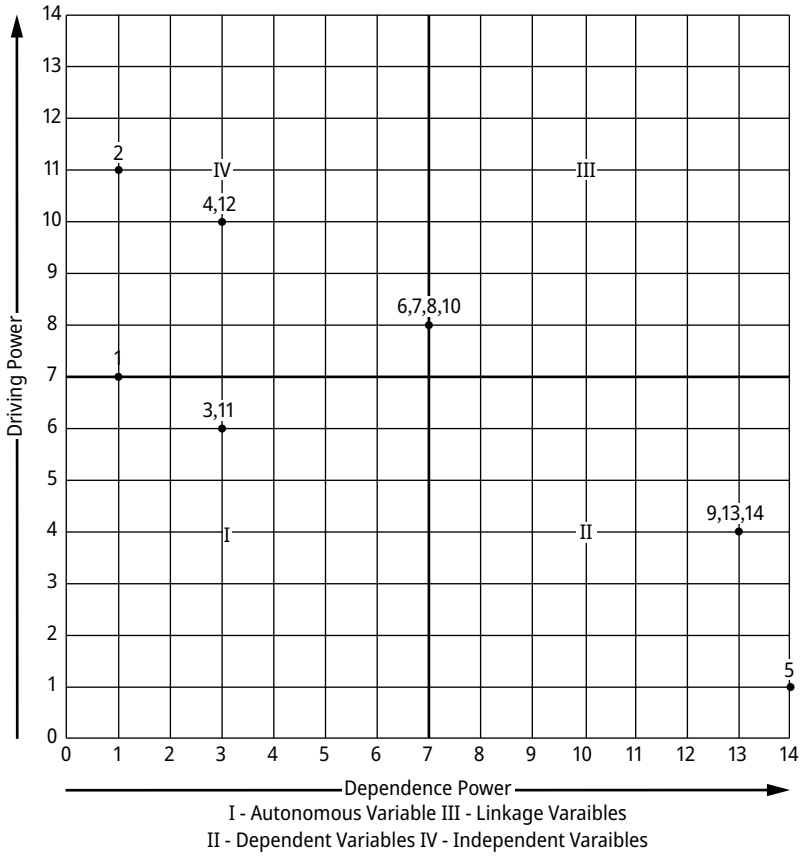


Figure 11.3: MICMAC Analysis.

influence higher-level issues. At the top (Levels 1 and 2), barriers like Complexity of AI Models (B5), Scalability Challenges (B9), Lack of Real-Time Capabilities (B13), and Integration with Legacy Systems (B14) have high dependence power but low driving power, making them dependence barriers. These will naturally improve as foundational and linkage-level barriers are addressed, indicating their reactive nature within the system. This structured hierarchy is crucial for stakeholders, helping prioritize efforts starting from the root causes to achieve effective and sustainable integration of AI in RM.

Before AI can fully enhance operational strength and manage risks, it must become a reliable resource. This results from the obstacles in AI-powered risk identification, assessment, mitigation, monitoring, and reporting. The ISM study intended the model to illustrate the linkage and impact of each barrier. This helped identify the most pressing obstacles at each level.

11.4.1 Risk identification phase

The key concerns that made it difficult to recognize possible hazards posed by AI during the risk identification phase were the absence of universal standards (B1) and the high cost of putting solutions into place (B2). Both of these issues came about as a result of the complexity of the situation. Even though it has a low reliance power of 1 and a pushing power of 7, B1 significantly impacts the restrictions of AI adoption. This indicates that it is less dependent on other hurdles.

High Implementation Expenditures (B2), with a driving power of 11, suggests that the organization must manage the early costs associated with AI technology to anticipate potential challenges. According to the findings of this study, applying AI might be difficult for SMEs and other organizations with few resources. Introducing new solutions is relatively costly, which is the reason behind this situation. This finding aligns with previous studies that indicate SMEs and organizations with limited resources are hesitant to adopt AI due to the substantial upfront costs required for implementation (Schönberger, 2023).

The Data Quality Issues (B3) barrier impedes risk detection. This barrier has a driving power of six and a reliance power of three. The fact that all of these characteristics are present with one another is where this originated. A reliable and accurate data set is essential for identifying possible threats by AI systems. AI models may produce false risk assessments should low-quality data be used as input. This point surfaced concerning the assessment of hazards. As a result, this shift reduces the efficacy of AI.

11.4.2 Risk assessment phase

As the investigation revealed, insufficient stakeholder assistance (B4) is a significant obstacle to risk assessment. To assess the challanges associated with technology, it is essential to have the support of an organization. Organizations are cautious about deploying advances since evaluations of AI are typically ambiguous and open to skepticism. One strategy for addressing people's unfavorable attitudes towards AI is to educate them about the numerous advantages of technology, such as its capacity to improve decision-making. Given the complex nature of AI algorithms (B5), risk assessment has become increasingly challenging. There is a significant cause for concern over the lack of comprehension of AI models (B7) owing to the complexity and ethical concerns (B6) linked with these models. AI models, particularly those built on deep learning, may be intricate for individuals lacking technical expertise. This dilemma has been further complicated by the ethical dilemmas and transparency challenges that beset AI. An expanding body of research (Machlev et al., 2022) underscores the importance of explainable AI (XAI) in resolving these challenges.

11.4.3 Risk mitigation phase

The project's risk mitigation phase uncovered a serious concern: a lack of competent workers (B8). This barrier has a dependability of seven and a driving power of eight. According to (Kinkel et al., 2022), data science, AI, and machine learning are all considerably understaffed. Recruiting and retaining workers with AI system development and maintenance expertise is a significant business challenge. Key personnel are required to guarantee that risk-reduction measures are implemented effectively. Due to a trained labor shortage, businesses cannot control risks using AI. Significant challenges were found during the risk mitigation phase, including scalability problems (B9) and trouble adhering to regulatory requirements (B10). Multiple researchers claim that AI systems struggle to adapt their abilities to various activities and environments. This limits their integration across large organizations. The use of AI in RM is constrained by the requirement that companies adhere to industry-specific data security regulations and procedures. This is especially crucial in several industries, such as banking and healthcare.

11.4.4 Risk monitoring & reporting phase

Many issues plagued the risk monitoring and reporting phase; however, cybersecurity threats (B12) and data protection (B11) were the most critical. This illustrates AI systems' need to acquire real-time data for constant risk assessment while adhering to privacy standards. With Driving Force 10, the Cybersecurity Risks (B12) showed how easy it is to hack AI systems, which makes them less helpful in controlling risks. Two more issues must be fixed before AI can be used for risk tracking and reporting. The first is that integration with older systems (B14) can't be used, and the second is that jobs can't be done in real time (B13). Because of bugs, using AI tools with creative technologies on older systems can be hard. Some AI systems can't analyze data in real time, making it hard for businesses to control risks correctly. These problems are caused mainly by outside factors that need to be recognized and dealt with in a planned way.

11.5 Discussion

The substantial implementation costs (B2), stakeholder engagement (B4), and Cybersecurity Risks (B12) deter many organizations from using AI in RM. These criteria are given the utmost attention. The substantial expense of AI is a considerable barrier for SMEs, depriving AI-driven RM solutions. Researchers (Enshassi et al., 2025) indicated that financial incentives or governmental support are necessary to promote the use of

AI. This aligns with the findings of the numerous research studies in the manufacturing context. The widespread use of AI is obstructed by many substantial hurdles, making the issue even more complex. These issues include, for example, the enigma of AI models (B7) and ethical dilemmas (B6). Given these issues, research papers on ethical, transparent, and successful AI systems have gained increased relevance.

In AI-driven RM, addressing these concerns will help establish trust among stakeholders, ultimately ensuring that management is fair and consistent. It is important to note that the lack of competent staff (B8) is a big problem. Consequently, this highlights the need for businesses to assist with projects to enhance their data science and AI capacities, employing the acquisition of new skills and training. Organizations cannot overcome these challenges without a meticulously designed strategy that includes investments in infrastructure, staff development, and comprehensive automation.

11.6 Conclusion

Companies must address issues at all levels to properly embrace AI in RM. Many systematic approaches exist to address several issues, including ethical issues, technological limitations, financial constraints, strategic investments, partner involvement, and legal compliance. Organizations must address the aforementioned issues before AI can completely realize its potential to alter how RM is done in all domains. Further studies should investigate how AI could help companies forecast risks and combine risk management techniques with emerging technologies like blockchain and the Internet of Things, which could improve RM practices. This method will enable more innovative and more flexible RM methods in manufacturing.

References

AlMahasneh, R., Hollósi, G., Ficzere, D., Bancsics, M., Lukovszki, C., & Varga, P. (2024). Uncovering Common AI Challenges Across Industrial Domains in the Transition to Industry 5.0. *2024 20th International Conference on Network and Service Management (CNSM)*, 1–7. https://doi.org/10.23919/CNSM62983.2024.10814318

Ångström, R. C., Björn, M., Dahlander, L., Mähring, M., & Wallin, M. W. (2023). Getting AI Implementation Right: Insights from a Global Survey. *California Management Review*, *66*(1), 5–22. https://doi.org/10.1177/00081256231190430

Bag, S., Sahu, A. K., Kilbourn, P., Pisa, N., Dhamija, P., & Sahu, A. K. (2022). Modeling barriers of digital manufacturing in a circular economy for enhancing sustainability. *International Journal of Productivity and Performance Management*, *71*(3), 833–869. https://doi.org/10.1108/IJPPM-12-2020-0637

Barbu, M., Vevera, A.-V., & Barbu, D.-C. (2024). *Standardization and Interoperability—Key Elements of Digital Transformation BT – Digital Transformation: Technology, Tools, and Studies* (L.-I. Cioca, L. Ivascu, F. G. Filip, & B. Doina (eds.); pp. 87–94). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-55952-5_5

Baryannis, G., Validi, S., Dani, S., & Antoniou, G. (2019). Supply chain risk management and artificial intelligence: state of the art and future research directions. *International Journal of Production Research*, *57*(7), 2179–2202. https://doi.org/10.1080/00207543.2018.1530476

Ben Ayed, R., & Hanana, M. (2021). Artificial Intelligence to Improve the Food and Agriculture Sector. *Journal of Food Quality*, *2021*(1), 5584754. https://doi.org/https://doi.org/10.1155/2021/5584754

Carvajal-Flores, D. F., Abril-Jiménez, P., Buhid, E., Fico, G., & Cabrera Umpiérrez, M. F. (2024). Enhancing Industrial Digitalisation through an Adaptable Component for Bridging Semantic Interoperability Gaps. In *Applied Sciences* (Vol. 14, Issue 6). https://doi.org/10.3390/app14062309

Cheng, B., Lin, H., & Kong, Y. (2023). Challenge or hindrance? How and when organizational artificial intelligence adoption influences employee job crafting. *Journal of Business Research*, *164*, 113987. https://doi.org/https://doi.org/10.1016/j.jbusres.2023.113987

Cheong, B. C. (2024). Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, *6*. https://doi.org/10.3389/fhumd.2024.1421273

Deiva Ganesh, A., & Kalpana, P. (2022). Future of artificial intelligence and its influence on supply chain risk management – A systematic review. *Computers & Industrial Engineering*, *169*, 108206. https://doi.org/https://doi.org/10.1016/j.cie.2022.108206

Edjlali, R. (2025). *Lack of AI-Ready Data Puts AI Projects at Risk*. Gartner. https://www.gartner.com/en/newsroom/press-releases/2025-02-26-lack-of-ai-ready-data-puts-ai-projects-at-risk

Engemann, K. J., & Miller, H. E. (2024). Toward revealing concealed risks for agile IT service management practices. *Information Systems and E-Business Management*. https://doi.org/10.1007/s10257-023-00666-8

Enshassi, M., Nathan, R. J., Soekmawati, & Ismail, H. (2025). Unveiling Barriers and Drivers of AI Adoption for Digital Marketing in Malaysian SMEs. *Journal of Open Innovation: Technology, Market, and Complexity*, 100519. https://doi.org/https://doi.org/10.1016/j.joitmc.2025.100519

Fan, Z., Yan, Z., & Wen, S. (2023). Deep Learning and Artificial Intelligence in Sustainability: A Review of SDGs, Renewable Energy, and Environmental Health. In *Sustainability* (Vol. 15, Issue 18). https://doi.org/10.3390/su151813493

Fehr, J., Citro, B., Malpani, R., Lippert, C., & Madai, V. I. (2024). A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. *Frontiers in Digital Health*, *6*. https://doi.org/10.3389/fdgth.2024.1267290

Ferrara, E. (2024). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. In *Sci* (Vol. 6, Issue 1). https://doi.org/10.3390/sci6010003

Golda, A., Mekonen, K., Pandey, A., Singh, A., Hassija, V., Chamola, V., & Sikdar, B. (2024). Privacy and Security Concerns in Generative AI: A Comprehensive Survey. *IEEE Access*, *12*, 48126–48144. https://doi.org/10.1109/ACCESS.2024.3381611

Govindan, K. (2022). Tunneling the barriers of blockchain technology in remanufacturing for achieving sustainable development goals: A circular manufacturing perspective. *Business Strategy and the Environment*, *31*(8), 3769–3785. https://doi.org/https://doi.org/10.1002/bse.3031

Gurjar, K., Jangra, A., Baber, H., Islam, M., & Sheikh, S. A. (2024). An Analytical Review on the Impact of Artificial Intelligence on the Business Industry: Applications, Trends, and Challenges. *IEEE Engineering Management Review*, *52*(2), 84–102. https://doi.org/10.1109/EMR.2024.3355973

Hangl, J., Krause, S., & Behrens, V. J. (2023). Drivers, barriers and social considerations for AI adoption in SCM. *Technology in Society*, *74*, 102299. https://doi.org/https://doi.org/10.1016/j.techsoc.2023.102299

Kinkel, S., Baumgartner, M., & Cherubini, E. (2022). Prerequisites for the adoption of AI technologies in manufacturing – Evidence from a worldwide sample of manufacturing companies. *Technovation*, *110*, 102375. https://doi.org/https://doi.org/10.1016/j.technovation.2021.102375

Kusiak, A. (2017). Smart manufacturing must embrace big data. *Nature*, *544*(7648), 23–25. https://doi.org/10.1038/544023a

Lada, S., Chekima, B., Karim, M. R. A., Fabeil, N. F., Ayub, M. S., Amirul, S. M., Ansar, R., Bouteraa, M., Fook, L. M., & Zaki, H. O. (2023). Determining factors related to artificial intelligence (AI) adoption among Malaysia's small and medium-sized businesses. *Journal of Open Innovation: Technology, Market, and Complexity*, *9*(4), 100144. https://doi.org/https://doi.org/10.1016/j.joitmc.2023.100144

Lee, E., Seo, Y.-D., Oh, S.-R., & Kim, Y.-G. (2021). A Survey on Standards for Interoperability and Security in the Internet of Things. *IEEE Communications Surveys & Tutorials*, *23*(2), 1020–1047. https://doi.org/10.1109/COMST.2021.3067354

Machlev, R., Heistrene, L., Perl, M., Levy, K. Y., Belikov, J., Mannor, S., & Levron, Y. (2022). Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, *9*, 100169. https://doi.org/https://doi.org/10.1016/j.egyai.2022.100169

Mazhar, T., Irfan, H. M., Haq, I., Ullah, I., Ashraf, M., Shloul, T. A., Ghadi, Y. Y., Imran, & Elkamchouchi, D. H. (2023). Analysis of Challenges and Solutions of IoT in Smart Grids Using AI and Machine Learning Techniques: A Review. In *Electronics* (Vol. 12, Issue 1). https://doi.org/10.3390/electronics12010242

Mennella, C., Maniscalco, U., De Pietro, G., & Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon*, *10*(4). https://doi.org/10.1016/j.heliyon.2024.e26297

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2023). Operationalising AI ethics: barriers, enablers and next steps. *AI & SOCIETY*, *38*(1), 411–423. https://doi.org/10.1007/s00146-021-01308-8

Nair, M., Svedberg, P., Larsson, I., & Nygren, J. M. (2024). A comprehensive overview of barriers and strategies for AI implementation in healthcare: Mixed-method design. *PLOS ONE*, *19*(8), e0305949. https://doi.org/10.1371/journal.pone.0305949

Oldemeyer, L., Jede, A., & Teuteberg, F. (2024). Investigation of artificial intelligence in SMEs: a systematic review of the state of the art and the main implementation challenges. *Management Review Quarterly*. https://doi.org/10.1007/s11301-024-00405-4

Peretz-Andersson, E., Tabares, S., Mikalef, P., & Parida, V. (2024). Artificial intelligence implementation in manufacturing SMEs: A resource orchestration approach. *International Journal of Information Management*, *77*, 102781. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2024.102781

Plathottam, S. J., Rzonca, A., Lakhnori, R., & Iloeje, C. O. (2023). A review of artificial intelligence applications in manufacturing operations. *Journal of Advanced Manufacturing and Processing*, *5*(3), e10159. https://doi.org/https://doi.org/10.1002/amp2.10159

Rezaei, J., Kothadiya, O., Tavasszy, L., & Kroesen, M. (2018). Quality assessment of airline baggage handling systems using SERVQUAL and BWM. *Tourism Management*, *66*, 85–93. https://doi.org/https://doi.org/10.1016/j.tourman.2017.11.009

Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: barriers and pathways forward. *International Affairs*, *100*(3), 1275–1286. https://doi.org/10.1093/ia/iiae073

Satvik, Naik, B. K., Dwivedi, R., & Yoshikuni, A. C. (2025). Modelling the blockchain technology adoption barriers in the Indian education management system. *Journal of Modelling in Management*, *ahead-of-p* (ahead-of-print). https://doi.org/10.1108/JM2-10-2024-0331

Schönberger, M. (2023). Artificial intelligence for small and medium-sized enterprises: Identifying key applications and challenges. *Journal of Business Management*, *21*, 89–112. https://doi.org/10.32025/JBM23004

Sharma, S. (2024). Benefits or concerns of AI: A multistakeholder responsibility. *Futures*, *157*, 103328. https://doi.org/https://doi.org/10.1016/j.futures.2024.103328

Singh Patyal, V., Kumar, R., & Singh Kushwah, S. (2021). Modeling barriers to the adoption of electric vehicles: An Indian perspective. *Energy*, *237*. https://doi.org/10.1016/j.energy.2021.121554

Singla, A., Sukharevsky, Alexander Yee, L., Chui, M., & Hall, B. (2025). *The state of AI: How organizations are rewiring to capture value*. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#/

Sinha, A., Sapra, D., Sinwar, D., Singh, V., & Raghuwanshi, G. (2024). Assessing and Mitigating Bias in Artificial Intelligence: A Review. *Recent Advances in Computer Science and Communications*, *17*(1), 1–10. https://doi.org/https://doi.org/10.2174/2666255816666230523114425

Sinha, S., & Lee, Y. M. (2024). Challenges with developing and deploying AI models and applications in industrial systems. *Discover Artificial Intelligence*, *4*(1), 55. https://doi.org/10.1007/s44163-024-00151-2

Surange, V. G., & Bokade, S. U. (2024). A Comparative Study of the Severity Ranking of Risks in the Indian Automobile Manufacturing Supply Chain Using PROMETHEE, VIKOR and TOPSIS. *Vikalpa*, *49*(4), 327–348. https://doi.org/10.1177/02560909241302201

Surange, V. G., Bokade, S. U., Singh, A. K., & Teli, S. N. (2021). Prioritization of roadblocks to adoption of industry 4.0 technologies in manufacturing industries using VIKOR. *Materials Today: Proceedings*, *50*, 2194–2200. https://doi.org/10.1016/j.matpr.2021.09.448

Surange, V. G., Sangode, P. B., & Gaikwad, L. M. (2025). Examining Obstacles to Industry 4.0 Adoption in Indian Manufacturing SMEs through ISM and TOPSIS Analysis. *Journal of The Institution of Engineers (India): Series C*, *106*(1), 213–225. https://doi.org/10.1007/s40032-024-01142-z

Surange, V. G., Suthar, J., Teli, S. N., & Sutrisno, A. (2024). Key Enablers for Transitioning to Circular Supply Chains in Electronics: An ISM MICMAC Analysis. *Jordan Journal of Mechanical and Industrial Engineering*, *18*(4), 823–834. https://doi.org/10.59038/jjmie/180414

Tariq, M. U., Poulin, M., & Abonamah, A. A. (2021). Achieving Operational Excellence Through Artificial Intelligence: Driving Forces and Barriers. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.686624

Umer, M. A., Belay, E. G., & Gouveia, L. B. (2024). Leveraging Artificial Intelligence and Provenance Blockchain Framework to Mitigate Risks in Cloud Manufacturing in Industry 4.0. In *Electronics* (Vol. 13, Issue 3). https://doi.org/10.3390/electronics13030660

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology*, *34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Warfield, J. N. (1974). Developing Interconnection Matrices in Structural Modeling. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-4*(1), 81–87. https://doi.org/10.1109/TSMC.1974.5408524

Whang, S. E., Roh, Y., Song, H., & Lee, J.-G. (2023). Data collection and quality challenges in deep learning: a data-centric AI perspective. *The VLDB Journal*, *32*(4), 791–813. https://doi.org/10.1007/s00778-022-00775-9

Zhang, Z., Song, X., Liu, L., Yin, J., Wang, Y., & Lan, D. (2021). Recent Advances in Blockchain and Artificial Intelligence Integration: Feasibility Analysis, Research Issues, Applications, Challenges, and Future Work. *Security and Communication Networks*, *2021*(1), 9991535. https://doi.org/https://doi.org/10.1155/2021/9991535

Natasha Blazheska-Tabakovska and Snezana Savoska

12 Artificial intelligence impact on education: Opportunities and risks

12.1 Introduction

Digital transformation, including artificial intelligence (AI) adoption, has become a key catalyst for growth, significantly enhancing quality of life. The impact of digitalization spans across global sectors, fostering innovation, inclusivity, and sustainability (Dignum, 2021). Education, as one of the most critical sectors of society, has experienced profound changes due to this technological shift.

Artificial intelligence has evolved from a futuristic idea to an indispensable part of daily life, with education at the forefront of this transformation. Educational leaders and institutions are increasingly acknowledging the transformative potential of artificial intelligence (AI) in education.

The integration of pedagogical, social, and technical affordances, augmented by AI capabilities, transforms playful learning environments into dynamic spaces conducive to effective and personalized learning experiences (Romero, Lameras, & Arnab, 2024).

Artificial Intelligence, in the educational context, refers to the application of machine learning algorithms and data analytics to enhance various aspects of the learning process. It encompasses a spectrum of applications, from intelligent tutoring systems and personalized learning platforms to smart content creation and automated grading systems. Generative AI (GenAI) can be defined as a "technology that (i) leverages deep learning models to (ii) generate human-like content (e.g., images, words) in response to (iii) complex and varied prompts (e.g., languages, instructions, questions)" (Lim, Gunasekara, Pallant, Pallant, & Pechenkina, 2023).

AI technologies are revolutionizing the academic landscape, unlocking unprecedented opportunities for personalized learning, administrative efficiency, and improved educational outcomes. AI has the potential to address some of the biggest challenges in education today while also innovating teaching and learning practices. AI is reshaping traditional educational models by offering personalized learning experiences tailored to individual student needs, making education more effective and engaging. AI-enabled technologies also assist in administrative tasks, streamlining operations and allowing educators to focus more on teaching. Additionally, AI provides valuable insights through data analytics, improving both the learning experience and overall educational efficiency.

Moreover, AI is accelerating progress toward Sustainable Development Goal 4 (SDG4), which aims to ensure inclusive, equitable, and quality education for all (UNESCO, Artificial intelligence in education, 2024).

https://doi.org/10.1515/9783111595306-012

AI has been widely applied to various educational technology platforms such as 1) Virtual Mentor, 2) Voice Assistant, for example, Google Assistant (Google), Siri (Apple), and Cortana (Microsoft). 3) Smart Content, 4) Presentation Translator. 5) Global Courses, for example, MOOCs, Udemy, Google AI, Alison, Khan Academy, edX, Udacity, Coursera, etc. 6) Automatic Assessment, 7) Personalized Learning for examples Ruangguru, etc. 7) Educational games, 8) Intelligent Tutoring System (ITS) or Intelligent Computer- Aided Instruction (ICAI) (Fitria, 2021).

As AI continues to evolve, the potential for further innovation in education is increasing. Teachers and school administrators are already using AI-powered tools to enhance their work, save time and improve their working experience, and the possibilities for AI's future role in education are immense. From customizing lessons to automating administrative workflows, AI is transforming education, making it more adaptive, efficient, and accessible for all. Wanga and al. identified four primary categories of AI applications in Education (AIEd), each with its corresponding subcategories: (1) Adaptive learning and personalized tutoring, (2) Intelligent assessment and management, (3) Profiling and prediction, and (4) Emerging technologies or products (Wanga, et al., 2024).

Despite its potential to advance education by enabling scalable, cost-effective, and impactful solutions, the integration of AI in education brings system-level risks and concerns about future implications. One significant risk is the potential for increased surveillance of students, raising ethical and privacy issues. There is also anxiety among educators about being replaced by AI. Algorithmic bias is a critical concern. Additionally, the lack of transparency in some AI applications, particularly those that operate invisibly, can erode trust and accountability. The risk of AI providing information that appears credible but is inaccurate or baseless further complicates its adoption. I also introduce risks related to data privacy and security, as well as "algorithmic discrimination," where automated systems may inadvertently create unfair learning opportunities or resource allocations for specific student populations [7].

12.1.1 AI for personalized learning: Unlocking individual potential

In the dynamic landscape of education, the global learning market, especially the e-learning market, has undergone a profound transformation. e-Learning has made a great leap from traditional methods to embrace innovative approaches. The year 2021 is just the beginning of a trajectory set for an exponential climb in growth in the sector. At the heart of this educational revolution lies the concept of personalized learning, a paradigm that changes educational experiences to the unique needs of individual learners. Studies of student learning have long shown that learning can be more effective if it is performed through private tutoring rather than teaching in a classroom. In classroom teaching, all the students have to listen to the same lectures regardless of the knowledge that they have and their learning preferences (Hasanu &

Blazheska-Tabakovska, 2018). Personalization of the learning process departure from the one-size-fits-all model, is driven by the recognition that students possess different learning styles, paces and preferences. The significance of personalized learning lies in its ability to cater to individual student needs.

Personalization of the e-learning system improves the quality of the system itself and allows learners to learn in a style that most suits them (Klašnja-Milićević, Vesin, Ivanović, & Z., 2011), (Blazheska-Tabakovska, Ivanović, Klašnja-Milićević, & Ivković, 2017). By recognizing and adapting to the diverse ways in which students learn, personalized learning fosters increased engagement, deeper understanding, and more meaningful educational outcomes. As students navigate their educational journeys, tailored experiences enhance their strengths, address their weaknesses, and empower them to thrive in a personalized learning environment.

The essence of personalized learning is to unlock individual potential. The teacher can't deliver a lecture that is customized to each student's existing knowledge. Personalized learning addresses these challenges. It is an adaptive and student-centric pedagogical approach that recognizes the individuality of each learner. It goes beyond the confines of a standardized curriculum, acknowledging that students possess distinct strengths, weaknesses, and learning styles. In personalized learning, an educational technique concentrates on individual students' learning processes and caters to their wants, talents, and concerns. One of the pivotal aspects of personalized learning is the nuanced tailoring of educational content to individual strengths, weaknesses, and learning styles. Rather than imposing a uniform pace on all students, personalized learning allows them to progress at a rate that aligns with their understanding. Personalized learning lets each person's wants and learning goals be met by changing things like the speed at which they learn, the materials they use, the order in which they learn them, the technologies they use, the quality of the materials, the way they are taught, and the materials they use to learn (Meehirr, 2023). It promises to offer them varied learning experiences to help them acquire knowledge. This method goes beyond earlier education, and people support the view that students could attain better learning when the teaching environment is customized (Claned, 2024).

AI for personalized learning has huge potential. AI can be utilized to personalize education in many different ways. The role of AI is not confined to merely automating tasks; it acts as a catalyst for a personalized and adaptive learning experience. Machine learning algorithms analyze vast datasets to identify patterns in student behaviour, learning styles, and comprehension levels. This allows educators to tailor their teaching methods and address the unique needs of each student. Furthermore, AI facilitates early identification of learning difficulties and provides timely interventions, ensuring a proactive approach to educational support.

Adaptive learning opportunities provide personalized learning plans tailored to each learner's needs. This approach allows learners to skip concepts they already know, receive targeted recommendations, and master all necessary concepts at their own pace. AI can identify patterns that might indicate the student's preferred learning

style or areas they struggle in (Chen, Zhang, & Zheng, 2018). AI can assist in crafting educational content that is personalized for each individual student. That content can be presented through different formats, e.g., text-based materials, videos, interactive simulations etc. AI-powered educational tools allow teachers to detect the areas where the student might face difficulties and proactively to provide additional support like practice exercises, remedial materials, or one-on-one tutoring. AI predictive analytics assess students' performance and have good reasoning in advance regarding problem areas. The research conducted by Khan and Jian showed that students in the AI-driven personalized learning group showed a statistically significant improvement in their grades compared to those in the traditional learning group and the students were more consistent in their performance, suggesting that AI tools provided them a steadier learning curve. Moreover, the students felt the AI tools made learning more engaging and catered to their individual learning needs (Khan & Jian, 2023).

AI can automate the assessment process and provide timely, personalized feedback to students. Intelligent grading systems can evaluate assignments, quizzes, and exams, enabling faster feedback delivery. This helps students identify their areas of improvement and provides targeted suggestions for further learning.

Also, AI can overcome the barriers to education and provide an individualized experience for students who have various types of learning disabilities and needs. AI can give extra aid to students with disabilities. For students with visual challenges, object recognition and text-to-speech technologies provide greater autonomy in accessing educational content. These tools convert visual information into formats that are more easily processed, enabling independent learning. With the integration of AI, the latest text-to-speech and screen reader technologies offer enhanced accuracy and context-sensitive readings, even for complex materials. This empowers students to engage with and understand their learning resources on their own. AI-driven tools also offer real-time transcription and closed captioning for students with hearing differences, ensuring they can fully participate in lectures and discussions alongside their peers. These adaptive technologies level the playing field, allowing all students to access, comprehend, and interact with educational content without relying on external assistance (Lam, 2024).

The transformative power of Artificial Intelligence in personalizing learning is undeniable, as demonstrated by its ability to identify students' learning styles, generate highly tailored educational content, anticipate and address potential learning challenges, facilitate intelligent and adaptive assessment, and provide customized support for students with diverse disabilities and specific individual needs (Figure 12.1). Its applications extend far beyond these examples, continuously reshaping the educational landscape in countless ways.

Additionally, Yang et al. in their study, emphasize the role of AI gamified platforms in personalizing game scenarios to enhance learning outcomes. Such platforms can make learning not just individualized but also engaging (Yang, Sinha, Adamson, & Rosé, 2018).

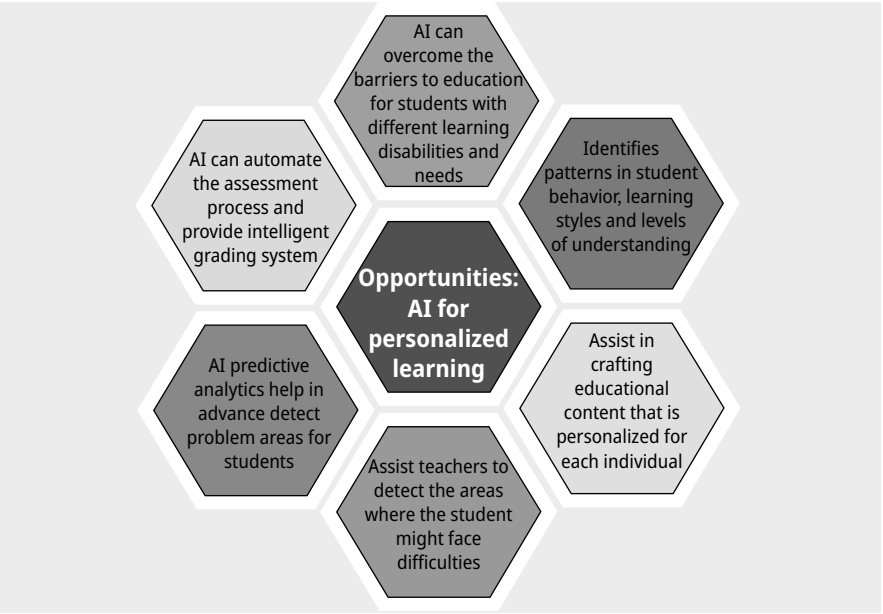


Figure 12.1: Opportunities: AI for personalized learning.

The possibilities of AIEd technologies for personalized approaches in education are great. The possibilities can be systematized in several areas: Social Networks, Expert Systems, Intellectual mentors and agents, Machine Learning Technology, Professional Educational Systems and Environments and Virtual Educational Environments (Tapalova & Zhiyenbayeva, 2022).

12.1.2 Benefits of implementing AIEd for personalised learning

The implementation of Artificial Intelligence in Education (AIEd) offers numerous benefits to all participants in the educational process: students, educators and educational institutions (Tapalova & Zhiyenbayeva, 2022).

The students benefit in many ways. Students gain the opportunity to learn in an enriched, adaptive learning environment equipped with personalized materials and metacognitive prompts. This enables them to take an active role in managing their own learning process, fostering responsibility in setting and achieving educational goals. AIEd allows students to learn at their own pace, providing a comprehensive educational experience that includes enhanced access to diverse disciplines and essential 21st-century job skills. Additionally, it equips students with practical professional competencies necessary for success in the workforce. By integrating AIEd, the students are empowered to take responsibility for their personal, professional, and aca-

demic development, while also gaining the ability to apply their acquired knowledge and skills in real-world contexts.

Educators benefit significantly from the integration of AI in education, gaining powerful tools to enhance teaching and learning processes. They can personalize learning experiences by analyzing both qualitative and quantitative learning data, tailoring educational strategies to meet individual student needs. AI enables the delivery of engaging and immersive educational programs in virtual environments, fostering a more interactive learning experience. Educators can generate relevant and practical learning content, developing individualized learning pathways that consider each student's strengths, weaknesses, talents, and challenges. Additionally, they can implement social and emotional learning programs remotely, ensuring holistic development even in virtual settings. AI allows for the automation and tracking of student progress in mastering professional skills, providing valuable insights into their development. It also helps educators identify the most effective teaching methods by considering the learning context and the student's current knowledge level. Finally, AI empowers educators to provide immediate, personalized feedback, enhancing the learning process and ensuring that students receive the support they need to succeed.

Lastly, but equally important, educational institutions also benefit significantly from the implementation of Artificial Intelligence in Education (AIEd) for personalized learning. Institutions gain the ability to instantly analyze large-scale training data for administrative purposes, enabling data-driven decision-making and efficient resource allocation. AIEd facilitates the implementation of more effective and streamlined educational processes, improving overall institutional efficiency. Moreover, AIEd offers the opportunity to enhance the quality of education by tailoring learning experiences to the unique characteristics and expectations of each student. By considering factors such as personality, talents, goals, and background, institutions can create a more inclusive and student-centered learning environment that fosters success and satisfaction for all learners.

12.2 ChatGPT

In November 2022, OpenAI, a lab that studies artificial intelligence, came out with a chatbot called ChatGPT (Pre-trained Transformer). ChatGPT (Chat Generative Pre-Trained Transformer) the artificial intelligence chatbot provided by OpenAI, attracted the world's attention in a short period. There are over 100 million users registered to explore the usage of ChatGPT in different fields, including education. ChatGPT is a conversational artificial intelligence interface that uses natural language processing (NLP), which interacts in a realistic way and even "answers follow-up questions, admits its mistakes, challenges incorrect premises, and rejects inappropriate requests" (OpenAI., 2023). While ChatGPT's primary function was to mimic human conversation,

its capabilities extend far beyond that; it can literally create new things, such as a poem, story, or novel, or act like anything within its capabilities. Although ChatGPT in the beginning has been labelled as a charlatan or a sophisticated parrot, as it does not truly understand what it is talking about (Alonso, n.d.), continuous improvements are being released or resolved in future versions or through integration with other plugins that enhance the versatility and capabilities of ChatGPT. Today, the powerful capacity of ChatGPT has surprised educators, teachers, and universities with its ability to generate quick and human-like responses.

The intelligent chatbots and virtual assistants can interact with students in a conversational manner (Janssen, Grützner, & Breitner, 2021). These tools can answer questions, provide explanations, offer feedback, and engage in dialogue to enhance the learning experience (Meehirr, 2023). With the chatbot possibility, students can recommend online study materials to help them complete their assignments more quickly. They can register themselves by completing online forms through the assistance tool, and they can also obtain admission details from universities.

Some educators and teachers believe that ChatGPT has a positive impact on learning. Teachers and students can use ChatGPT to save time and improve learning. It can help teachers to save time in several ways, such as generating sample responses and feedback for creating banks of questions and providing different levels of examples of writing when setting assessments. For students, ChatGPT can help them learn in a more efficient way. Students can ask ChatGPT to provide suggestions on specific subjects, and ChatGPT can explain how its suggestions can improve the analysis and argument. ChatGPT not only helps teachers and students to prepare and complete their teaching and learning but also assists principals in pinpointing areas where intervention is needed, especially detecting signs of stress, anxiety, or depression in students. This helps school management to identify the needs and provide early intervention in a more efficient way (Henebery, 2023).

The User experience, according to experiences of using ChatGPT and identify potential concerns that might be faced when using it in education, points to a more global projective goal: not just attain effectiveness, efficiency, and satisfaction, but it aims to enhance the entire experience of the user, from the expectation, through interaction and finally the reflection about the experience (Beccari & Oliveira, 2011).

12.2.1 Enhancing teaching and learning: The benefits of chatgpt

ChatGPT stands out among other chatbots due to its advanced capabilities, versatility, and adaptability, making it a powerful tool in modern education (Duong, Da, & Hanh Nguyen, 2024). It has the potential to significantly enhance traditional teaching methods and transform the overall learning experience. One of its key strengths is its ability to provide personalized learning experiences. ChatGPT tailors its responses based on individual learners' proficiency, learning style, language, and pace, offering tar-

geted assistance that helps students focus on their specific areas of weakness while reinforcing their strengths.

Additionally, ChatGPT serves as a reliable source of foundational knowledge, simplifying complex concepts and making them accessible to learners of all levels. It acts as a supplementary tutor, answering questions and clarifying doubts outside of traditional classroom hours. For educators, ChatGPT significantly reduces workload by automating repetitive tasks such as grading assignments, providing feedback, and creating lesson plans. It also supports administrative functions like drafting emails, scheduling, and maintaining student records, allowing teachers to concentrate on more impactful teaching activities.

ChatGPT's 24/7 availability ensures that students can access help whenever needed, regardless of time or location. This makes it especially beneficial for distance learning programs, students in different time zones, or individuals who prefer learning at their own pace. It offers immediate access to information and resources, enhancing the overall learning experience. Moreover, ChatGPT excels in generating creative ideas for projects, essays, and research topics, and breaks down complex subjects into manageable parts, facilitating better comprehension.

ChatGPT also assists students in drafting essays, reports, and assignments, ensuring coherence and clarity while suggesting improvements in grammar, vocabulary, and style. This contributes to improved learning outcomes, as personalized support and immediate feedback foster a deeper understanding and better retention of knowledge, leading to enhanced academic performance. Students can revisit topics multiple times, building mastery and confidence in their abilities.

The tool is inclusive and scalable, supporting learners with diverse needs, including those with disabilities, by providing accessible educational resources. It scales effortlessly, serving large numbers of students without compromising the quality of interaction. ChatGPT's multifaceted capabilities make it a game-changer in the educational landscape. By blending personalized learning, efficient task management, and innovative teaching aids, it empowers both educators and students to achieve their goals more effectively. As technology continues to evolve, the integration of tools like ChatGPT will play a pivotal role in shaping the future of teaching and learning.

In addition to these benefits, ChatGPT can serve as an additional resource alongside traditional course material. Students can request further explanations, examples, or alternative perspectives, enabling them to explore topics in greater depth. This can improve understanding, reinforce learning, and encourage independent research. ChatGPT is also particularly useful for language learners, offering practice opportunities, language correction, and assistance with vocabulary and grammar. It can engage students in conversational exchanges, helping them enhance their communication skills and express their thoughts and ideas more effectively.

For instructors and teaching assistants, ChatGPT can handle routine questions, provide quick references, and offer guidance on frequently asked questions, freeing up their time to focus on higher-level tasks such as facilitating discussions, conducting

critical analysis, and providing personalized feedback to students. Furthermore, integrating ChatGPT into the curriculum can introduce innovative and interactive learning experiences, such as simulations or virtual role-playing scenarios. These experiences can improve student motivation, critical thinking, and creativity, making learning more engaging and effective.

Overall, ChatGPT's ability to support personalized learning, enhance teaching efficiency, and foster innovation in educational experiences makes it an invaluable tool in modern education. Its adaptability and scalability ensure that it can meet the diverse needs of both students and educators, ultimately transforming the way education is delivered and experienced.

ChatGPT can support research efforts by assisting with literature reviews, data analysis, and generating hypotheses. Researchers can leverage the model's ability to sift through vast amounts of information and provide insights on complex topics. ChatGPT can aid in discovering relevant sources, summarizing findings, and even suggesting potential research directions (Michel-Villarreal R., Vilalta-Perdomo, Salinas-Navarro, Thierry-Aguilera, & Gerardou, 2023).

12.3 Concerns

Artificial intelligence (AI) can transform education by improving teaching and learning outcomes.

But the rapid evolution of Artificial Intelligence (AI) also creates a myriad of new challenges (Dwivedi, et al., 2023). There is a critical need for the careful implementation of AI in educational settings, ensuring a balance between the advantages of technology and the preservation of ethical standards and pedagogical integrity.

The Beijing Consensus on Artificial Intelligence aims to address the opportunities and challenges presented by AI concerning education and provides 44 recommendations grouped into different aspects that help understand the magnitude of the task (UNESCO, Beijing Consensus on Artificial Intelligence and Education. I, 2019). Some of these aspects include planning AI in educational policies; AI for the management and delivery of education; AI to support teaching and teachers; AI for learning and assessment of learning; developing values and skills for life and work in the AI era; AI for providing lifelong learning opportunities for all; promoting equitable and inclusive use of AI in education; gender-responsive AI and AI for gender equality etc. All these 44 aspects bring a different challenge.

One of the primary challenges of incorporating AI in education is ensuring its ethical use. The large amount of data required for AI algorithms to effectively personalize learning can raise questions about student privacy and data security (Zhou & & Brown, 2015). It is crucial to establish clear guidelines and regulations to protect students' personal information and ensure that AI algorithms are transparent and unbi-

ased (Stefanic, 2024). Any discussion surrounding privacy or transparency must be approached with utmost caution. Systems engineers, developers, and administrators responsible for creating and deploying AI in educational institutions should involve legal counsel early in the design process. AI and ML algorithms must be thoroughly assessed for potential vulnerabilities and liabilities, with careful consideration given to both cybersecurity risks and user experience concerns (Schaeffer, Coombs, Luckett, Marin, & Olson, 2024). Data confidentiality refers to safeguarding information within a system or application to prevent unauthorized access. Without secure, trusted data from connected sensors, devices, systems, and applications, these components become vulnerable to cyber manipulation, which can undermine the integrity of AI decision-making. Moreover, the "black box" nature of certain AI algorithms might make it difficult for educators to understand or trust the educational recommendations being made. Additionally, there are availability risks associated with AI applications, as generative AI is not always accurate or reliable, which can further compromise its effectiveness in educational settings (Pavlou, 2023).

A key concern in the context of personalized learning, as well as the learning process in general, is the validation of educational resources. In traditional learning, educational resources are typically validated based on key criteria such as content accuracy, age appropriateness, pedagogical relevance, and cultural and social sensitivity. This validation process often takes a year or longer. However, in most cases today, AI tools require no such validation (Giannini, 2024).

Another primary concern revolves around ChatGPT and similar generative AI tools. AI Large Language Model (LLM) systems, such as ChatGPT, empower students with the ability to create content based on prompts, a process known as prompt engineering (Baidoo-Anu & Owusu Ansah, 2023). These ideas can be used in whole or in part by students to submit as their work, which raises questions about authorship (Schaeffer, Coombs, Luckett, Marin, & Olson, 2024). Who deserves credit for generating the idea—the student or the AI? Globally, educators are skeptical and worried that students may use ChatGPT to cheat on assignments, thereby undermining the integrity of learning assessments, certifications, and qualifications (Anders, 2023), (Chris Gilliard, 2023). Some teachers and schools thought the AI chatbot did not help students develop their professional skills, like critical thinking and problem-solving. The integration of ChatGPT into academic settings raises concerns about academic integrity, plagiarism detection, and the increased risk of cheating. Additionally, there are concerns about its potential impact on students' critical thinking skills, their ability to evaluate information, analyze complex concepts, and solve problems independently. Adaptive learning platforms are not exempt from these issues. In those circumstances, while some educational institutions banned the use of ChatGPT (Christopher Harris, 2023), others cautiously welcomed the arrival of GenAI (Cannix Yau, 2023). Tlili et al. suggest that teachers should embrace the new AI technology rather than ban it (Tlili A., Bozkurt, Hickey, Huang, & Agyemang, 2023). Many schools and universities have adopted a more progressive approach of support. They believe that instead of prohib-

iting the use of (Baidoo-Anu & Owusu Ansah, 2023) (Baidoo-Anu & Owusu Ansah, 2023)AI, students and staff should be supported in using these tools effectively, ethically, and transparently (Russell Group, 2023). In this context, there is a clear need for teachers to receive training in the effective implementation of AI in the classroom. Additionally, teachers and educational institutions must offer clearer guidance and support to help students understand how to integrate ChatGPT into the learning environment.

The challenge of cheating is closely tied to assessment systems—how can we accurately assess learning outcomes? Tests that were once considered 'unhackable' are now vulnerable to manipulation by AI applications. Furthermore, if the data used to train AI systems contains biases, as Barker and Kelsey (2007), bias occurs when one value in a sample space is more likely to be chosen than another. These biases can be learned and perpetuated by the AI, potentially affecting the assessment of learning outcomes. Concerns about bias in evaluating learning outcomes are legitimate. At the same time, there are ongoing concerns about the content of learning materials, which often lack thorough reviews for racial, political, or other types of bias. However, these issues have not been extensively explored.

One of the significant risks in implementing AI technologies in education is that they may not be accessible to all students, regardless of their background or abilities. The cost of AI-powered tools, such as augmented reality, avatars, and voice synthesis, remains a major barrier, limiting their widespread adoption and potentially widening the educational gap for those who cannot afford such technologies, either at an individual or institutional level. This includes considerations for students with disabilities, those from low-income families, and those without access to reliable internet connections. It is essential to design AI-powered educational tools and platforms that are inclusive and provide equal opportunities for all students.

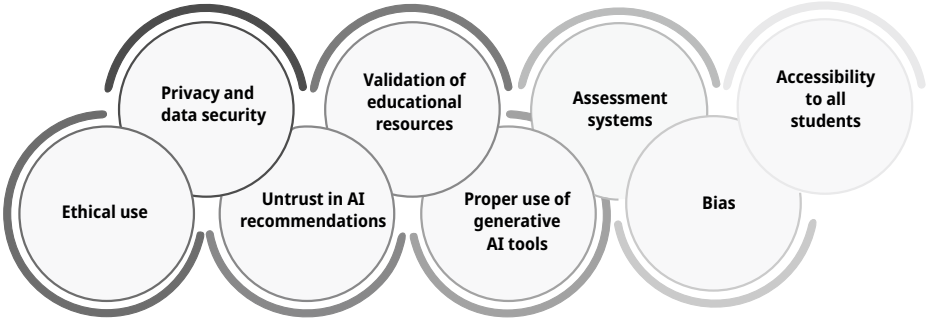


Figure 12.2: Challenges: AI in education.

Figure 12.2 highlights the key challenges and considerations surrounding the use of AI in education, including ethical implementation, data privacy, building trust, validating educational resources, ensuring the responsible use of generative tools, address-

ing biases in assessment systems, and promoting accessibility for all learners, among many other emerging concerns.

With the emergence of AI tools like ChatGPT, there is a potential risk that traditional methods of knowledge-sharing in education could decline. In the past, students and educators engaged in collaborative learning environments, asking questions, sharing expertise, and building a collective knowledge base. However, as AI models can generate answers almost instantly, there is concern that the number of people engaging in knowledge-sharing may decrease. This shift could lead to a reduction in the desire to share expertise. AI models, such as ChatGPT, do not inherently accumulate knowledge; they rely on continuous input and training. If new information isn't consistently fed into these systems, their development could stagnate, making them less useful for educational purposes and ultimately hindering the overall growth of collaborative learning and critical thinking in the educational process (Evans, 2024).

12.4 Conclusions

The future of AI in education and workforce training holds great potential. As AI continues to evolve, it is evident that we will see an increased integration of AI technologies in the education sector, which can address many important challenges in education. AI has the potential to revolutionize traditional teaching methods by providing tailored learning experiences that cater to individual needs and preferences. Advancements in AI technology will further shape the educational landscape, offering more innovative tools and approaches to personalized learning.

Personalized learning, supported by AI, will empower students to control how and when they learn, allowing them to choose the learning methods that suit them best and progress at their own pace. AI maximizes learning outcomes by customizing content and pace according to each student's specific needs. This approach ensures that students are presented with appropriate challenges and support. Additionally, using AI for targeted learning helps teachers quickly assess what students already know and identify areas they need to focus on.

Machine learning and predictive analytics offer insights into individual learning progress, preferences, and challenges, enabling continuous adaptation of learning strategies. By embracing AI, educators can enhance the quality of education, deliver more effective training, and empower learners to reach their full potential.

Adaptive learning platforms will become increasingly sophisticated, analyzing data on student performance to identify both areas of strength and areas that require further attention.

To fully realize the potential of AI in education, it is essential to be aware of the risks associated with its use and understand how to address these challenges effectively.

In response to one of the main concerns – the potential for students to use AI tools to cheat on assessments or otherwise bypass the learning process, some educators have called for rethinking traditional assessments and shifting towards more creative and open-ended tasks that machines cannot quickly solve (Kulkarni, Kotturi, Bernstein, & Klemmer, 2015). Also, a response to this risk is the approach to design tasks that require creativity, critical thinking, and problem-solving skills that machines cannot replicate. One example is project-based learning tasks, where students must work collaboratively and solve complex problems (Ayala-Pazmiño, 2023).

As a response to another challenge related to the use of AI in education—widening the digital divide and exacerbating existing inequalities in educational opportunities and outcomes—systematic efforts are required from both local and state authorities. Together with educators, these authorities must ensure that AI tools are accessible to all students and take proactive steps to bridge the digital divide, such as providing access to technology and internet resources in underserved communities.

The National Education Association (NEA) emphasizes the importance of addressing the ethical and privacy concerns related to the use of AI in education (National Education Association, 2020). This challenge also requires heightened awareness and vigilance regarding data security and algorithmic bias

The teachers must be trained in integrating AI into their teaching practices, evaluating AI tools for their effectiveness, and identifying and addressing potential ethical issues. By providing appropriate training to educators, we can ensure that AI is used responsibly and effectively in the classroom and that students benefit most from these powerful tools.

Not all educators share the same perspective on the use of AI, particularly tools like ChatGPT. Some argue that teachers have a responsibility to prepare students for the future, and they question whether relying on AI could undermine the effectiveness and integrity of teaching and learning.

In their exploration of AI's impact on education, Popenici and Kerr highlighted a key concern: AI tools may eventually replace certain aspects of teachers' roles. They emphasized the need for educators to develop a new pedagogical philosophy that aligns with the demands of the 21st century (Popenici & Kerr, 2017).

While AI offers immense potential to personalize education, it is essential to recognize that human teachers remain irreplaceable. They play a crucial role in fostering supportive, engaging, and motivating learning environments. The integration of AI with human expertise will not replace teachers, but rather enhance their ability to provide a more personalized and effective curriculum, ultimately benefiting every student

References

Alonso, C. (n.d.). *Ojo con ChatGPT, que es un charlatán mentirosillo!*. El futuro está por hackear. Zenda. Retrieved October 2024, from https://bit.ly/44dEbCk

Anders, B. A. (2023). Is using ChatGPT cheating, plagiarism, both, neither, or forward thinking? *Patterns*, *4*(1), 100694. https://doi.org/10.1016/j.patter.2023.100694

Ayala-Pazmiño, M. (2023). Artificial intelligence in education: Exploring the potential benefits and risks. *593 Digital Publisher CEIT*, *8*(3), 892–899. https://doi.org/10.33386/593dp.2023.3.1827

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *SSRN*. https://doi.org/10.2139/ssrn.4394114

Barker, E., & Kelsey, J. (2007). *Recommendation for random number generation using deterministic random bit generators* (NIST Special Publication 800-90). U.S. Department of Commerce, National Institute of Standards and Technology.

Beccari, M. N., & Oliveira, T. L. (2011). A philosophical approach about user experience methodology. In *International Conference of Design, User Experience, and Usability* (pp. 13–22). Springer.

Blazheska-Tabakovska, N., Ivanović, M., Klašnja-Milićević, A., & Ivković, J. (2017). Comparison of e-learning personalization systems: Protus and PLeMSys. *International Journal of Emerging Technologies in Learning (iJET)*, *12*(1), 57–70.

Cannix Yau, K. C. (2023). University of Hong Kong temporarily bans students from using ChatGPT, other AI-based tools for coursework. *South China Morning Post*. ``Retrieved from https://www.scmp.com/news/hong-kong/education/article/3210650/university-hong-kong-temporarily-bans-students-using-chatgpt

Chen, L., Zhang, D., & Zheng, X. (2018). Personalized learning resource recommendation algorithm based on hybrid filtering for online education. *Educational Technology Research and Development*, *66*(1), 1–23.

Chris Gilliard, P. R. (2023). You're not going to like how colleges respond to ChatGPT. *Slate*. Retrieved October 2024, from https://slate.com/technology/2023/02/chat-gpt-cheating-college-ai-detection.html

Christopher Harris, A. T. (2023). Can you tell between a year 6 student and AI? Teachers say they can. *The Sydney Morning Herald*. Retrieved October 2024, from https://www.smh.com.au/national/nsw/can-you-tell-between-a-year-6-student-and-ai-teachers-say-they-can-20230120-p5ce5s.html

Claned. (2024). *The role of AI in personalized learning*. Retrieved October 2024, from https://claned.com/the-role-of-ai-in-personalized-learning/

Dignum, V. (2021). The role and challenges of education for responsible AI. *London Review of Education*, *19*(1), 1–11. https://doi.org/10.14324/LRE.19.1.14

Duong, Thi, Thuy, Mai; Da, Can.; Hanh Nguyen (2024). The use of ChatGPT in teaching and learning: A systematic review through SWOT analysis approach. *Frontiers in Education*, *9*, 1328769. https://doi.org/10.3389/feduc.2024.1328769

Dwivedi, Y., Kshetri, N., Hughes, L., Slade, E., Jeyaraj, A., Kar, A., . . . & Sharma, S. K. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2022.102642

Evans, H. (2024). What's the catch: Risks and concerns of using AI in education. *Velvetech Blog*. Retrieved February 2025, from https://www.velvetech.com/blog/ai-in-education-risks-and-concerns/

Fitria, T. N. (2021). Artificial intelligence (AI) in education: Using AI tools for teaching and learning process. *Prosiding Seminar Nasional and Call for Paper STIE AAS*, *1*(1), 47–52.

Giannini, S. (2024). Generation AI: Navigating the opportunities and risks of artificial intelligence in education. *UNESCO*. Retrieved from https://www.unesco.org/en/articles/generation-ai-navigating-opportunities-and-risks-artificial-intelligence-education

Hasanu, M., & Blazheska-Tabakovska, N. (2018). Designing intelligent tutoring system based on Bayesian network. In *International Conference on Applied Internet and Information Technologies* (pp. 69–72).

Henebery, B. (2023). How principals can leverage AI to improve staff and student wellbeing. *The Educator*. Retrieved October 2024, from https://www.theeducatoronline.com/k12/news/how-principals-can-leverage-ai-to-improve-staff-and-student-wellbeing/282013

Janssen, A., Grützner, L., & Breitner, M. H. (2021). Why do chatbots fail? A critical success factors analysis. In *Forty-Second International Conference on Information Systems (ICIS)*.

Khan, M. J., & Jian, O. (2023). Personalized learning through AI. *Advances in Engineering Innovation*, *5*(1), 47–52.

Klašnja-Milićević, A., Vesin, B., Ivanović, M., & Budimac, Z. (2011). E-learning personalization based on hybrid recommendation strategy and learning style identification. *Computers & Education*, *56*(3), 885–899.

Kulkarni, C. C., Kotturi, Y., Bernstein, M. S., & Klemmer, S. R. (2015). Peer and self-assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *22*(2), 1–31.

Lam, H. A. (2024). AI for accessibility – Accelerate inclusion: The remarkable impact of AI on accessibility in education. *RMIT Vietnam*. Retrieved February 2025, from https://www.rmit.edu.vn/students/student-news-and-events/student-news/ai-impact-accessibility-education

Lim, W., Gunasekara, A., Pallant, J., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, *21*, 100790. https://doi.org/10.1016/j.ijme.2023.100790

Meehirr, K. (2023). How AI is personalizing education for every student. *eLearning Industry*. Retrieved October 2024, from https://elearningindustry.com/how-ai-is-personalizing-education-for-every-student

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and opportunities of generative AI for higher education as explained by ChatGPT. *Education Sciences*, *13*(9), 1–18. https://doi.org/10.3390/educsci13090856

National Education Association. (2020). *Preparing educators for AI*. https://www.nea.org/resource-library/preparing-educators-ai

OpenAI. (2023). *ChatGPT: Optimizing language models for dialogue*. Retrieved October 2024, from https://openai.com/blog/chatgpt/

Pavlou, C. (2023). Availability bias in the AI era: Understanding its impact on decision-making. *TalentLMS Blog*. Retrieved February 2025, from https://www.talentlms.com/blog/availability-bias-ai-era/

Popenici, S. A., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, *12*(1), 1–13.

Rainie, L. (2014). The six types of Twitter conversations. *Pew Research Center*.http://www.pewresearch.org/fact-tank/2014/02/20/the-six-types-of-twitter-conversati

Romero, M., Lameras, P., & Arnab, S. (2024). Affordances for AI-enhanced digital game-based learning. In *Creative Applications of Artificial Intelligence in Education* (pp. 117–128). Palgrave Macmillan.

Russell Group. (2023). *Russell Group principles on the use of generative AI tools in education*. https://russellgroup.ac.uk/media/6137/rg_ai_principles-final.pdf

Schaeffer, D., Coombs, L., Luckett, J., Marin, M., & Olson, P. (2024). Risks of AI applications used in higher education. *The Electronic Journal of e-Learning*, *22*(6), 60–65. https://doi.org/10.34190/ejel.22.6.3457

Stefanic, D. (2024). *AI and the future of personalized education*. Retrieved October 2024, from https://hyperspace.mv/future-of-personalized-education/

Tapalova, O., & Zhiyenbayeva, N. (2022). Artificial intelligence in education: AIEd for personalised learning pathways. *The Electronic Journal of e-Learning*, *20*(5), 639–653.

Tlili, A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, *10*, 15. https://doi.org/10.1186/s40561-023-00248-1

UNESCO. (2019). *Beijing Consensus on Artificial Intelligence and Education*. International Conference on Artificial Intelligence and Education. https://bit.ly/3n7wBIK

UNESCO. (2024). *Artificial intelligence in education*. Retrieved January 2025, from https://www.unesco.org/en/digital-education/artificial-intelligence

U.S. Department of Education. (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations*. Washington, DC: Office of Educational Technology.

Wanga, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems With Applications*, *252*, 1–19. https://doi.org/10.1016/j.eswa.2023.120223

Yang, D., Sinha, T., Adamson, D., & Rosé, C. P. (2018). Turn on, tune in, drop out: Anticipating student dropouts in Massive Open Online Courses. *Proceedings of the 2013 NIPS Data-Driven Education Workshop*, *11*(15), 14–21.

Zhou, L., & Brown, D. (2015). The ethical challenges of ubiquitous personalized learning environments in higher education: A constructivist perspective. *Ethics and Information Technology*, *17*(4), 283–293.

Contributors

Natasha Blazheska-Tabakovska
University of St. Kliment Ohridski – Bitola, North Macedonia

Gregory Bramble
IBM, USA

Krista N. Engemann
Accenture, USA

Kurt J. Engemann
Iona University, USA

Lokpriya M. Gaikwad
SIES Graduate School of Technology, India

Muhammad Z. Hasan
University of Central Punjab, Pakistan

Muhammad Z. Hussain
Bahria University, Pakistan

Matthew LaGana
MBIA, USA

C. Christopher Lee
Central Connecticut State University, USA

Rachelle Loyear
ASIS International, USA

Dron M. Mandhana
Villanova University, USA

Holmes E. Miller
Muhlenberg College, USA

Robert Richardson
Iona University, USA

Jeanne Sheehan
Iona University, USA

Heechang Shin
Iona University, USA

Muhammad R. Siddique
MCB Bank, Pakistan

Ore Soluade
Iona University, USA

Snezana Savoska
University of St. Kliment Ohridski – Bitola, North Macedonia

Vinod G. Surange
Symbiosis International (Deemed University), India

Yusuf J. Ugras
La Salle University, USA

https://doi.org/10.1515/9783111595306-013

Index

accessibility 7, 24, 42, 110, 116, 165, 188, 197, 244
accounting xi, 165–184
adaptive malware vii, 60
adversarial attacks vi–vii, 38, 63, 77, 79–84, 87, 92
adversarial machine Learning 63
AI agents 182–183
AI automation 90, 172
AI capabilities viii, 18, 42, 106, 109–121, 184, 233
AI defense 16, 21, 35–36, 38–39, 59–60, 68–69, 72, 74, 78, 84–87, 171, 176
AI evaluation 109
AI governance 29, 48–50, 54, 65, 89, 92, 174, 176, 213
AI integration vii, 52–53, 74–75, 174, 209, 211, 213
AI model bias vi–vii, 43, 50, 82, 85, 215
AI model complexity xii, 18, 44, 60, 72, 78, 144, 153, 156, 165, 209, 212, 215, 219, 225
AI research 95
AI risk v, 21–22, 24–28, 43, 45, 47–49, 62–63, 87, 91, 131, 174, 184, 212–213
AI training data 28, 119
AI transparency 118, 176
AI-driven network security vii, 77–92
AI-powered deception technologies 79, 92
AI-powered phishing vii, 60
algorithmic bias x, xii, 27, 50, 234, 245
anomaly detection vii, 39, 41, 45, 60, 64, 77–78, 84, 86, 89–90, 111, 183
Artificial Intelligence (AI). v–xiii, 3, 14, 35–36, 44, 47, 77, 95, 109, 127, 137, 149, 152, 169, 172–173, 175, 180, 184, 209, 233, 241
Artificial Superintelligence (ASI) 131
automation v, vii, xi, 5, 10, 13, 20, 22, 24–25, 27, 36–37, 39, 41, 43, 50–53, 59, 68, 74, 77–80, 90–92, 109, 114–115, 169–173, 175–177, 182–183, 187–188, 194–195, 217, 228, 238
automobiles xi, 7–8, 12, 195
autonomous vehicles xi, 198, 205

behavioral analysis 39, 60
beneficial use 130–131, 137–142
bias in AI vi, 12, 48, 62–63, 68, 215
bridging frontiers 97
buses xi, 188–189, 191, 195

ChatGPT 96–98, 100–101, 107–108, 116, 238–245
computer vision 112, 116
concentrated capital ownership vi, 6, 11, 30
continuous monitoring ix, 29, 40, 43, 45, 49, 71, 75, 83–84, 86–87, 89–90, 174, 216
critical barriers xii, 209
Cross-Impact Matrix Multiplication Applied to Classification (MICMAC) xii, 209
cyber resilience 59, 75
cybersecurity vi–vii, 13, 23, 25, 28, 36, 38, 53, 59–75, 77–79, 81–82, 84, 86–92, 111, 113, 168, 174, 176, 181, 184, 190, 197–198, 214, 216, 219, 223, 227, 242, 251

data breach 60, 150
data poisoning 36, 62, 73, 79–80, 85
data privacy xii, 25, 43, 62, 66–67, 81, 83, 87, 149–150, 153, 168, 174, 176, 213, 216, 219, 234, 243
deepfake vi, ix, 61, 79, 86, 127–144
detrimental use 130
digital transformation 233
disinformation ix, 35, 37, 131–132, 135, 137–138, 142, 144

economic inequality 4–7, 10, 16, 20
education xii–xiii, 7, 9, 27–29, 53, 61, 96, 199–200, 233–245
e-learning 234–235
employment v, 25, 62, 131, 172, 190, 193–194, 197
encryption vi, x, 28, 44, 77–78, 80, 150, 154–155, 160, 166, 174, 176, 181
engineering viii, 40, 61–62, 86, 96, 100, 103–106, 108, 159, 218, 242
Enterprise Security Risk Management (ESRM) vi, 35–55
ethical concerns 18, 47, 114, 212, 215, 223, 226
ethical dilemmas vii, xii, 13, 66–67, 226, 228
ethics in AI 29, 47–50, 53, 62, 66, 74–75, 79, 84–85, 87, 106, 118, 176, 184, 228, 234
explainability 19, 28, 49, 81, 83–84, 118–119, 174, 176
Explainable AI (XAI) 49, 80, 118, 179, 226

federated learning vii, 81, 88, 91, 153, 158–161
first-order risk ix, 129, 142
fraud detection 43, 63–64, 73–74, 80, 82, 84, 111, 113, 118, 166, 169–171, 173–176, 183–184

https://doi.org/10.1515/9783111595306-014

GDPR (General Data Protection Regulation) vii, 48, 65, 67, 80, 92, 149, 168
General AI (AGI) viii, 131, 182–183
Generative AI (GenAI) ix, xii, 13–14, 30, 35, 60, 127–144, 175, 233, 242
governance vi–vii, xi, 19, 28–29, 35, 48–50, 54, 65, 74, 77, 83, 88–89, 92, 95–96, 118, 130, 149, 169, 174, 176, 184, 211, 213
ground truth 105, 117–119

hallucination 95, 97, 119
Health Insurance Portability and Accountability Act (HIPAA) 43, 149, 153
healthcare vii, ix–x, 13, 15–17, 21, 28, 43, 46, 52, 62–63, 68, 70, 109, 118, 149–161, 180, 213, 227
human understanding viii, 96–97, 99–100, 102
human-AI collaboration in security 50–53
human-AI interaction 19, 115–117
human in the loop (HITL) 95–108

incident response vii, 35, 39, 45, 53, 64–65, 68–70, 75, 90
Indian manufacturing sector 217
individualized learning experience 236, 238
intelligent tutoring systems xii, 233–234
Internet of Things (IoT) 35, 38, 40, 43, 48, 54, 82, 88, 150, 158, 228
Interpretive Structural Modelling (ISM) xii, 209–228
Intrusion Detection System (IDS) 64, 70, 81

Large Language Models (LLMs) 37, 49, 97, 101–102, 105, 107, 113, 116, 242
legacy systems integration 214–216, 219, 225
Light Detection and Ranging (LiDAR) 112, 189–190

Machine Learning (ML) vii, xi, 17, 37–39, 41, 44, 46, 48, 53, 59–63, 65, 69–70, 72, 78–79, 86, 88–89, 99, 104–105, 109, 112–113, 115–116, 151–154, 158, 170, 173, 175, 178–179, 183, 189, 227, 233, 235, 237, 244
malicious GPT 60–61
malicious use 130–131, 138
misinformation ix, 8, 10, 37, 54, 116, 127–129, 137
misuse vi, ix, 49, 131, 137–142, 160, 174
model vulnerabilities 79, 82, 84

Narrow AI (ANI) 131
National Institute of Standards and Technology (NIST) vii, 27, 47–48, 67

organizational decision-making 23

personalized learning 233–242, 244
physical security 38–39, 54
policy x, 11, 27, 44, 66–67, 143, 156, 194
precautionary principle 28–29
predictive analytics x, 11, 38–39, 43, 47, 75, 78, 89, 91, 113, 171, 173, 175, 183–184, 236–237, 244
privacy attacks 62, 80, 91–92, 150, 156, 160
privacy-preserving AI 81, 87, 88
productization 103, 105

radar 71, 189–190, 195
real-time capabilities 79, 214, 216, 225
regulation vi–vii, xiii, 13, 20, 26–30, 43, 47–48, 65–67, 80, 92, 115, 149, 168, 175–176, 179, 193, 195, 198, 202–204, 227, 241
regulatory compliance xi, 28, 49–50, 161, 168–169, 173–174, 176, 179, 184
representative democracy 128, 130
responsible AI vii, 29, 118, 174, 176
risk analysis vi, 6, 21–27, 178, 211
risk framework v, 6–22
Risk Management (RM) v–xiii, 21, 24, 26–27, 35–55, 74, 85, 87, 91, 106, 121, 136, 169, 173–175, 177–178, 184, 209–228, 251
risk mitigation viii, 20, 24, 27–28, 86, 103–105, 142–144, 213, 216, 227
risk-based strategy vi, 141–143

safety ix, xi, xiii, 28–30, 37, 43, 115, 188, 190–192, 194, 198–199, 202–203, 212
scalability issues xii
second-order risk ix, 129, 142
self-driving xi, 63, 115, 191–194, 198
self-healing networks 91–92
self-healing systems 68
sensitive patient information 150, 155
stakeholder support 29, 209, 211, 223, 226
strategic AI adoption 176
surveillance vi, 12, 18–19, 25, 27, 35, 38–39, 43, 47–49, 80, 234

taxis xi, 189–191, 195, 199
technological advancement ix, 3, 24–25, 96, 99–100, 102, 168–169, 217
technological disruption 20, 24, 41, 43, 168
threat detection 35, 37–39, 44–45, 47, 49, 54–55, 59–60, 71, 75, 78–79, 84, 90–91, 111
threats vii, ix, 23, 26–27, 35–55, 59–60, 64–65, 67–71, 73–75, 77–92, 128–129, 136, 142, 150, 161, 166, 168, 172, 174, 176, 180, 184, 226–227
transparency in AI ix, 29, 174, 184, 212, 215, 223
trucks xi, 104, 189–192
Turing test 99

ultrasonic sensors 189–190

zero-trust security architectures vii

Developments in Managing and Exploiting Risk

The objective of this multi-volume set is to offer a balanced view to enable the reader to better appreciate risk as a counterpart to reward, and to understand how to holistically manage both elements of this duality. Crises can challenge any organization, and with a seemingly endless stream of disruptive and even catastrophic events taking place, there is an increasing emphasis on preparing for the worst. However, being focused on the negative aspects of risk, without considering the positive attributes, may be shortsighted. Playing it safe may not always be the best policy, because great benefits may be missed.

Analyzing risk is difficult, in part because it often entails events that have never occurred. Organizations, being mindful of undesirable potential events, are often keenly averse to risk to the detriment of capitalizing on its potential opportunities. Risk is usually perceived as a negative or downside, however, a commensurate weight should also be given to the potential rewards or upside, when evaluating new ventures. Even so, too much of a good thing may create unintended consequences of risk, which is also an undesirable situation. *Developments in Managing and Exploiting Risk* provides a professional and scholarly venue in the critical field of risk in business with emphasis on decision-making using a comprehensive and inclusive approach.

Vol. 1: Safety Risk Management: Integrating Economic and Safety Perspectives. Edited by Kurt J. Engemann and Eirik B. Abrahamsen

Vol. 2: Project Risk Management: Software Development and Risk. Edited by Kurt J. Engemann and Rory V. O'Connor

Vol. 3: Organizational Risk Management: Managing for Uncertainty and Ambiguity. Edited by Krista N. Engemann, Kurt J. Engemann and Cliff Scott

Vol. 4: Socio-Political Risk Management: Assessing and Managing Global Insecurity. Edited by Kurt J. Engemann, Cathryn F. Lavery and Jeanne Sheehan

Vol. 5: Cybersecurity Risk Management: Enhancing Leadership and Expertise. Edited by Kurt J. Engemann and Jason A. Witty

Vol. 6: Artificial Intelligence Risk Management: Ensuring Beneficial Outcomes. Edited by Kurt J. Engemann