Chapter

1

AI and AI Governance

THIS CHAPTER COVERS AIGP DOMAIN I, "UNDERSTANDING THE FOUNDATIONS OF AI GOVERNANCE," SUBDOMAIN I.A, "UNDERSTAND WHAT AI IS AND WHY IT NEEDS GOVERNANCE." THE FOLLOWING TOPICS ARE COVERED:

- ✔ Definitions and types of AI systems
- ✔ The types of risks and harms that AI systems may cause
- ✔ Why AI systems are different from other systems, and why governance is needed
- ✔ The principles of responsible AI

THE AI GOVERNANCE PROFESSIONAL (AIGP) PERFORMANCE INDICATORS IN THIS SUBDOMAIN INCLUDE:

- ✔ Know the generally accepted definitions and types of AI.
- ✔ Understand the differences in AI model types.
- ✔ Identify the types of risks and harms posed by AI to individuals, groups, organizations and society (e.g., misalignment with objectives, ethics and bias risk, and complexity and scalability).
- ✔ Identify the unique characteristics of AI that require a comprehensive approach to governance (for example, complexity, opacity, autonomy, speed and scale, potential for harm or misuse, data dependency, and probabilistic versus deterministic outputs).
- ✔ Identify and apply the common principles of responsible AI (for example, fairness, safety and reliability, privacy and security, transparency and explainability, accountability and human-centricity).

In all its forms, AI brings unprecedented capabilities to organizations in every business sector to improve efficiency and quality. But, like other emerging technologies, organizations are apt to rush into significant, impactful projects to bring about sweeping business transformation, even before they understand its capabilities and risks. This chapter includes a brief survey of the types of AI systems, with discussions on capabilities, risks, and reasons why organizations should develop a governance structure to ensure their AI systems deliver the results they expect.

The Types of AI

Artificial Intelligence (AI) is an interdisciplinary field that combines computer science, statistics, mathematics, and engineering to create systems capable of performing tasks that typically require human intelligence. The primary goal of AI is to design machines that can think, learn, and problem-solve autonomously.

This section includes definitions and examples of various types of AI.

Definition of AI

Artificial Intelligence (AI) is the simulation of human intelligence processes by machines, generally computer systems. These processes include learning, reasoning, problem-solving, perception, and language understanding.

Some AI systems are designed to perform specific tasks, such as image recognition and generation, speech processing, natural language understanding, decision-making, and more. For instance, tools like HeyGen create realistic videos of a specific subject from a photograph and a script.

Other AI systems are designed to perform a broad range of tasks, including large language models (LLMs) like Claude from Anthropic and Microsoft Copilot. For example, virtual assistants like Siri and Alexa use AI to interpret spoken language, understand requests, and provide relevant responses.

AI Capabilities

AI is often categorized into levels of capability, based on the extent of the system's ability to replicate human cognitive abilities. These are:

- *Narrow AI* (or *Weak AI*) refers to a class of AI systems designed and trained to perform a specific or narrow set of tasks. These systems do not possess general intelligence or consciousness and focus only on one function type. One example is IBM Deep Blue, an AI designed to play chess at an expert level. Another example is the *large language model (LLM)* AI (also known as *generative AI*), as implemented in ChatGPT, Claude, Grok, and others.
- An *AI agent* is an autonomous system that perceives its environment through incoming sensor information and then takes action to achieve specific goals.
- *General AI* (or Artificial General Intelligence—AGI) is a form of AI that can understand, learn, and apply intelligence across various tasks, much like a human being can. AGI is capable of solving nearly any problem it encounters, just as humans can. For example, an AI system capable of managing all or part of an organization, including performing tasks like strategic decision-making and customer service, and even creative tasks like writing or creating product or process designs.
- *Artificial Superintelligence (ASI)* (or *Superintelligent AI*) is an AI that surpasses human intelligence in all aspects—creativity, problem-solving, emotional understanding, and decision-making. While this level of AI does not yet exist, it is often discussed in AI research and ethical discussions, due to its potential risks and benefits.
- *God-like AI* is an AI at a level beyond Artificial Superintelligence (ASI), where AI capabilities exceed human cognition so profoundly that they become incomprehensible to humans across all domains.

As of this writing, no AGI or Superintelligent AI system has been disclosed to the general public.

AI Functionalities

AI systems are also categorized in terms of their functionality, based on the types of tasks or processes they can perform. These include:

- *Reactive machines* are AI systems designed to respond to specific stimuli or inputs. They do not retain memories of past interactions, nor do they learn from them. Reactive machines operate based on pre-programmed rules and cannot improve on their own. An example is IBM's Deep Blue chess player, which is trained to play chess at an expert level but does not remember past games.

- *Limited memory AI* systems with limited memory retain information for short periods and use that information to improve their performance. They learn from past experiences or historical data to enhance their decision-making capability. Still, because their memory is limited, they typically discard information when it is no longer relevant. For example, autonomous vehicles use limited memory AI to process traffic data, road conditions, and previous driving patterns. They use this information for navigation and decisions about speed and direction, but they soon discard information due to limited resources.
- *Reasoning AI systems* possess advanced abilities to analyze and correlate patterns and make informed predictions. Most common LLMs today fall in this category. They are even capable of passing standardized tests (for example, bar testing) with performance exceeding that of the average human.
- *Theory of mind AI* systems are a theoretical, advanced form of AI. This AI could understand and simulate human emotions, beliefs, intentions, and thoughts. In the future, an AI system could interact with humans in a manner similar to human relationships, including understanding and responding to emotions. A fictional example is the HAL-9000 computer in the movie *2001: A Space Odyssey*.
- *Self-aware AI* is the highest level of AI that possesses consciousness and self-awareness. Self-aware AI would be able to recognize itself in its environment and understand its own existence. Like superintelligent AI, self-aware AI is currently a theoretical concept. An example of self-aware AI would be a humanoid robot capable of introspection, understanding its own limitations, and even questioning its purpose.
- *Robotics* is a type of AI system capable of performing kinetic tasks, such as carrying and moving objects, and more advanced tasks, such as meal cooking.

AI Techniques

Another way to classify AI systems is by their techniques or methods to perform tasks. These methods help us understand how AI systems work. These include:

- *Machine learning (ML)* is a technique that enables systems to learn from data, improve their performance, and make decisions without explicit programming. In ML, *algorithms* are trained on large datasets and "learn" to make predictions or decisions based on patterns identified in the data. Spam filters, for example, use ML to analyze past emails and identify patterns (such as specific words or sender behaviors) to help classify new emails as spam.
- *Deep learning* is a subset of machine learning that uses artificial neural networks consisting of multiple layers. A deep learning system is designed to process large volumes of data, often unstructured, and is particularly effective at tasks such as speech and image recognition. A good example is Google Translate, which can translate spoken or written text into different languages.

- *Natural language processing (NLP)* systems can understand, interpret, and generate human language. NLP combines linguistics and machine learning to process text and speech, making interactions between humans and machines more intuitive. An example is ChatGPT, a chatbot that uses NLP to process and respond to user inputs in human language.
- *Supervised learning* is a type of AI system training that uses labeled data. For example, an ML-powered spam filter is trained with examples of spam messages and legitimate messages. In another example, a robot is trained with a dataset of images to later identify various types of objects.
- *Unsupervised learning* is a type of AI system training that uses unlabeled data. The AI system is expected to detect patterns on its own. For example, given a customer relationship database, an AI system is directed to classify it as it sees fit.
- *Semi-supervised learning* is a type of AI system training that uses a small amount of labeled data and a large amount of unlabeled data. The AI system first learns from the labeled data and is then trained on the unlabeled data to predict labels. Over time, performance will improve. For example, an organization is training an AI-powered image recognition system but only has a small collection of images to train. The AI system is trained on the labeled data and continues learning with the unlabeled data.
- *Reinforcement learning* is a type of machine learning where an AI agent learns through interaction and receives feedback in the form of rewards or penalties. The objective is for the AI system to maximize its rewards through trial and error. An example is a robot that learns to navigate a maze.
- *Expert systems* are AI systems that simulate the decision-making ability of a human expert in a particular field, using rules and knowledge bases to make inferences and solve complex problems. For example, a medical diagnostic system can analyze patient symptoms and medical history to suggest a diagnosis or treatment.
- *Zero-shot learning (ZSL)* is a machine learning use case in which an AI model is trained to recognize and categorize objects or concepts without having seen any examples of those categories or concepts beforehand.

Understanding these and other definitions is critical for navigating and governing AI. Because some AI governance professionals do not have deep technical knowledge, they must understand concepts like these to help them achieve better credibility with their technical peers, while striving toward better business outcomes.

Risks and Harms Posed by AI

AI has brought transformative changes across many industries by enhancing efficiency, creating new opportunities, and solving complex problems. However, as AI becomes more integrated into business and society, it introduces risks and potential harms. Understanding these risks is essential for professionals managing and governing AI systems.

Let's explore the various types of risks and harms that AI can pose to individuals, groups, organizations, and society. Each risk is explained with examples to provide clarity and context.

Risks and Harms to Individuals

The potential risks and harms to individuals take on two forms: when individuals interact directly with AI and when organizations use AI-driven processing and decision-making with information about individuals.

- **Privacy violations:** One of the most significant risks AI poses to individuals is the potential for privacy violations. AI systems, particularly those that rely on large datasets and *Retrieval Augmented Generation (RAG)* systems, can expose sensitive personal information, resulting in privacy breaches. An AI that has been trained on or has access to personal or *personally identifiable information (PII)* without anonymization can potentially leak sensitive data when prompted with related queries. For example, if an AI has been exposed to usernames, passwords, or email addresses, it might be able to reproduce this information in its responses. Similarly, it could attempt to predict a user's login details when asked. Another example can be facial recognition systems used by retail stores to track customer movements. These systems could collect biometric data without consent, leading to concerns about surveillance and personal privacy.
- **Bias and discrimination:** AI systems can perpetuate or even amplify existing societal biases, which can harm individuals, particularly those from marginalized groups. This occurs when AI models are improperly designed or trained on biased data. Recent examples include AI-powered decision-making systems that help screen and qualify employment candidates, loan applicants, and financial aid enrollees. Organizations are often unaware of the bias in their systems until victims speak out.
- **Cybersecurity risks:** Many IT organizations do a substandard job of protecting their systems and the sensitive data they store and process. This seems more common when organizations introduce new technologies like AI—they get excited about bringing in the new toy without adequate thought on protecting it. Individuals become victims of identity theft, financial loss, and the exposure of their private information. As organizations race to build and implement AI systems, fundamental security practices are often being overlooked. We are seeing organizations fail big; from training and fine-tuning AI on non-anonymized data to deploying it in systems that lack proper security controls and violate principles like least privilege and segregation of duties. Ultimately, it's the individual who pays the price, as their privacy and data security are often neglected.
- **Loss of visibility, autonomy, and control:** AI-powered decision-making and processing systems often take humans out of the loop, reducing a person's autonomy and ability to influence outcomes. Healthcare and financial services are two industries where people

feel like they are being treated like a number instead of a person. For instance, AI-driven diagnostic tools might override a healthcare provider's decision or misinterpret a patient's condition, causing the individual to lose control over their treatment plan and potentially leading to adverse health outcomes. This situation can occur further up the food chain when a health insurance company must approve a procedure or treatment before it can commence.

I'm just scratching the surface here. As AI proliferates into more organizations and performs more functions, I think it is likely that we will see more risks like these.

Risks and Harms to Groups

Beyond the impact on individuals, AI can also exact significant consequences for groups, such as communities, families, and social networks. AI poses risks to groups that tend to be more systemic, with far-reaching social implications.

- **Amplification of social inequality:** AI systems that operate at scale can exacerbate existing inequalities. Decision-making AI systems that derive profile data from social media can exhibit biases that affect specific groups. For example, an AI-powered loan qualification system might offer higher interest rates or deny loans to individuals in particular communities or groups.
- **Erosion of social cohesion:** AI-powered platforms may inadvertently create echo chambers with little diversity of ideas, favoring sensational content that increases polarization and reduces tolerance of viewpoints. For example, some social media platforms may prioritize emotionally charged content, creating more division and eroding societal trust.
- **Cybersecurity risks:** AI adoption is more widespread in critical infrastructure (transportation, utilities, healthcare) and business. As stated earlier, organizations are not great at securing new technologies like AI until security problems occur. For example, a large bank might implement an AI-driven chatbot that can be tricked into revealing information about other customers. Cybersecurity-related risks are discussed throughout this book.

As before, these are notable examples but not the only ones.

Risks and Harms to Organizations

Organizations are adopting AI systems to improve efficiencies, quality, and decision-making and to reduce certain operational costs. AI poses unique risks to the organizations using them, including:

- **Operational risks:** Organizations relying on AI for critical processes, including inventory management, fraud detection, and customer service, are vulnerable to disruptions if the

AI system fails or behaves unpredictably. The more integral the AI system, the greater the potential impact of failure. For example, an AI-driven retail store inventory system malfunction could result in too much or too little stock of some products, affecting customer satisfaction, revenue, and brand damage.

- **Legal and regulatory risks:** Laws and regulations concerning privacy and AI are emerging and changing. Organizations not paying attention to these developments are at risk of violating relevant regulations, some of which could lead to public sanctions and reputation damage. For example, an organization lacking effective data privacy governance can run afoul of privacy laws if its AI-enhanced systems process PII unlawfully. On the other hand, regulatory bodies worldwide are also struggling to keep up with the rapid development and deployment of AI systems.
- **Financial risks:** Organizations employing AI in resource management or decision-making systems may experience poor outcomes. Further, adopting new technologies such as AI can exceed cost estimates, affecting financial results. For instance, an organization using AI-driven securities trading could experience significant losses if the AI system misjudges market conditions and events.
- **Cybersecurity risks:** Organizations are increasingly adopting AI-powered cybersecurity tools to improve their defenses and reduce the likelihood or impact of cyberattacks. Like other types of AI-powered systems, it's critical that the organization understand how the AI-powered system is supposed to work, and do so consistently. For example, an organization purchased an AI-powered network intrusion prevention system, blocking network traffic from a critical business partner, resulting in an hours-long outage and negative publicity. Further, organizations must protect their AI systems from poisoning that could enable them to produce erroneous output.
- **Supply chain risks:** Organizations using vendor-supplied AI systems are at risk of attacks on those vendors, where the attackers insert malicious code into the software that is distributed to its customers. The 2020 SolarWinds attack is a textbook example of this risk.
- **Misalignment risks:** Due to limited AI literacy and a lack of understanding of how these models operate, behavior that appears benign in testing environments can lead to unintended consequences when deployed. For example, an AI tasked with maximizing productivity, measured by the number of items produced per worker, might conclude that reducing the number of workers is the most effective way to achieve that goal.
- **Reputation risks:** If realized, all the risks cited in this section can lead to public disclosure, resulting in public backlash and brand damage. For example, an AI-driven targeted ad system could select "tone-deaf" ads, leading to public outcry and reputation damage.

The risks and harms to organizations in this section can be considered from the organization's perspective, which includes the same risks to individuals and groups described in the earlier sections. For example, an organization using an AI-powered employment application screening tool can cause harm to the individuals unfairly treated, as well as to the organization itself, if the harm is publicized.

Risks and Harms to Society

The impact of AI on society is the most profound, as the risks and harms brought by AI can reshape social, economic, and political structures. These effects can transcend the scope of individual organizations and groups, affecting entire nations, regions, and global systems.

- **Job displacement:** The use of automation in organizations can lead to the displacement of workers in industries, including manufacturing, transportation, information technology, and customer service. Indeed, as I am writing this chapter, I see news articles portending decreased demand for software engineers, transportation, customer service representatives, and other jobs. Further, some organizations are no longer hiring entry-level personnel in some industries, leading to new college grads struggling to find employment. For example, the adoption of autonomous vehicles reduces the demand for drivers.
- **Economic inequality:** Biases in training data could result in unequal treatment of persons, resulting in or exacerbating economic inequality. For example, employment and loan applicants could be treated unequally.
- **Erosion of human autonomy and agency:** Society could become overly dependent on automated, AI-powered decision-making, further eroding human autonomy, when decisions about individuals, groups, and society could be made by algorithms rather than people. For instance, governments could increasingly rely on AI algorithms to make public policy decisions, rather than engaging with the electorate in traditional ways.
- **Ethical and moral challenges:** As AI systems become more powerful and are relied upon to make decisions, ethical dilemmas will result, such as when systems are tasked with making life-or-death decisions. These challenges become particularly pronounced when AI decisions conflict with societal values or norms. For example, an autonomous taxi with passengers could be faced with an unavoidable accident in which its passengers or nearby pedestrians may be injured or killed. How should the autonomous vehicle respond?
- **Cybersecurity and privacy:** The potential scope of cybersecurity and privacy issues is bounded only by the scope of the systems themselves. In the event of an AI-related malfunction, large-scale public utilities, control systems, and surveillance systems could inflict considerable harm, including bodily injury and death. For instance, a large electric power grid using AI decision-making to manage electrical loads could misinterpret incoming signals, resulting in widespread power failures.
- **Manipulation of public opinion:** Today's social media platforms use AI-driven algorithms to control information feeds to increase engagement and ad revenue. The potential societal and political power may be too compelling to resist. For example, contrary to the law, a national government could manipulate a popular social media platform to favor certain election candidates.

This section reviewed the potential impact of AI at the individual, group, organization, and societal levels. The risks discussed here are rarely isolated to just one level but will spill over and impact many more areas. For example, a large employer utilizing an AI-driven employment screening function to identify suitable candidates for interviews may have improperly trained its selection algorithm, resulting in harm to specific applicants, people groups, the organization itself, and potentially society (at least in the region(s) where the organization operates).

Understand me: I am not against AI. On the contrary, I am an optimist and believe AI presents incredible opportunities for individuals, groups, organizations, and society. But anywhere you find great opportunities, you also find significant risks. I believe that this cosmic balance is unavoidable.

For readers who are (or will be) responsible for AI governance, I hope this section has provided practical, high-level information to help you better understand the range and types of risks associated with AI. Later chapters in this book explore these and other risks in greater detail.

We've Seen This Mania Before

It is said that there is nothing new under the sun. When it comes to AI, with all the hype, mania, and promises to solve every problem, this is not the first time that a new technology has made otherwise rational people a little crazy. Stepping back in time, we had the advent of the personal computer in the 1980s, the Internet in the 1990s, the cloud in the 2000s, and now AI. Each case involved organizations rushing to implement these technologies, with little regard for rational thought and decision-making. Organizations can get drunk on the prospect of first-mover advantage. The dot-com crash of the early 2000s is evidence that organizations rushed to "go online" without first creating sound business plans and defining specific problems to be solved. With AI, it's the same but also different—it is embedded in everything today, and bolting on AI without carefully understanding the risks is fraught with more danger than in previous iterations of new, shiny technology.

Characteristics of AI Requiring Governance

AI systems have evolved rapidly, becoming integrated into every level of humanity, from everyday individual lives to organizations and national governments. As AI continues to transform governments and industries, its complexity and the potential consequences of its deployment necessitate the creation of governance structures to manage them and ensure favorable outcomes. Unlike traditional information systems, AI introduces new challenges

that require tailored governance strategies to ensure ethical use, minimize risks, and optimize benefits. With great power comes great responsibility. *Governance* provides the guardrails.

This section explores the unique characteristics of AI that necessitate a comprehensive approach to governance. These characteristics include:

- Complexity
- Opacity
- Autonomy
- Speed and scale
- Potential for harm and misuse
- Data dependency
- Privacy
- Intellectual property
- Probabilistic vs. deterministic outputs

Each of these is discussed in more detail in this section.

Complexity

AI systems are often incredibly complex regarding their underlying models and the complex environments in which they operate. This complexity requires governance frameworks that can address their technical and operational intricacies. Complexity arises from advanced techniques such as neural networks and deep learning, where the interrelationships between inputs and outputs are not easily understood.

For example, the healthcare industry uses AI-powered diagnostic tools with deep learning algorithms trained on vast amounts of medical data. These systems can identify patterns that humans might miss, and their complexity makes it difficult for healthcare providers to understand exactly *why* the system made a particular diagnosis. This raises concerns about explainability, interpretability, accountability, and transparency.

A governance structure must ensure that AI systems are explainable, that management is accountable for all outcomes, and that systems are auditable. This includes developing transparency standards and documentation practices, ensuring that AI systems can be monitored effectively to assess and mitigate risks.

Opacity

Some AI systems are, by their very nature, "black boxes," particularly those using machine learning and deep learning models. Essentially, it can be challenging to understand how inputs are transformed into outputs, making the decision-making process of the AI system opaque and mysterious. This opacity presents significant governance challenges, especially in high-stakes fields like finance, healthcare, and criminal justice, where decisions based on

AI predictions can have serious consequences. The opposite of opacity is *explainability*, in which humans can understand and explain how an AI system made a particular decision or created a specific output. Explainability is similar to *interpretability*, which is an AI system's ability to understand the relationships between patterns in data.

For example, a bank's AI-powered loan approval system might reject an application without clearly explaining the factors influencing its decision. If the rejection is based on biased data, the opacity of the AI system makes it difficult for an applicant to contest the decision or understand why it was made. Likewise, the bank may have difficulty explaining precisely why the loan was denied, potentially putting the bank in regulatory jeopardy.

A governance structure must ensure that certain AI systems incorporate explainability and *transparency* in their design and training. This includes requiring that AI systems provide clear, understandable reasons for decisions, especially in areas like finance and healthcare, where transparency can protect individuals' rights, prevent discrimination, and protect organizations against legal action.

Autonomy

Some AI systems are *autonomous*, meaning they make decisions or take actions without human initiation or intervention, increasing the complexity of managing AI systems, as it can be challenging to predict how an *AI agent* will behave in every situation. Autonomous systems may be able to evolve and adapt over time, based on new data, adding challenges to governance.

For example, an autonomous vehicle may make a split-second decision to avoid a collision by swerving, potentially putting pedestrians at risk. The vehicle's AI-driven decision-making process is autonomous and cannot be easily overridden or influenced by human operators (if present) during critical moments.

A governance structure should identify appropriate limits on the actions of AI systems, ensure safety measures are in place, and regularly monitor performance. The governance function must enact clear policies to define what AI systems can and cannot do and address and establish *accountability* for decisions made by autonomous systems.

Speed and Scale

AI systems process data and make decisions much faster than humans, enabling them to operate at scale across a wide range of applications. The speed at which AI systems operate is both a strength and a risk. AI's ability to process vast amounts of data at high speeds makes it possible to solve problems impossible for humans to address manually. However, the speed and scale of AI can also lead to unintended consequences (particularly electricity and computing costs) if systems are not properly governed.

For example, a social media platform's AI algorithm automatically recommends content to millions of users, optimizing for engagement and, ultimately, advertising revenue. However, this can also spread harmful misinformation before it is detected and addressed.

Governance structure must be prepared for the speed and scale and the resource impacts of AI decision-making, necessitating real-time monitoring, response protocols, and appropriate controls to prevent and mitigate large-scale risks such as misinformation, financial fraud, and security breaches.

The Global Impact of AI

According to *Scientific American*, data centers (including those running AI systems) accounted for about 1.5% of global electricity usage, and this is projected to double by 2030. At that time, AI will be consuming as much electricity as the entire country of Japan. In the United States, AI-driven data centers are expected to account for fully half of the increase in electricity usage by 2030.

Potential for Harm and Misuse

As discussed earlier in this chapter, AI systems have the potential to be used for harmful purposes and cause unintended negative consequences. Whether through malicious actors or unforeseen behavior, AI's potential for harm requires strong governance structures to detect and prevent abuse and manage risk. AI systems can be tested in the design phase to detect the potential for harm.

For example, AI-driven surveillance systems could be used by authoritarian governments to monitor citizens' activities, suppress dissent, and infringe upon privacy rights (this is not theoretical but is already the case in some countries). Similarly, AI-powered weapons in military applications could malfunction or be misused, leading to ethical and safety concerns.

Governance structures must establish safeguards to detect and prevent harmful uses of AI, including mechanisms for *ethical oversight*, *accountability*, and *security controls*. In some jurisdictions, such as the EU, regulatory frameworks define the permissible uses of AI and ensure that AI is not deployed in ways that violate human rights or present risks to public safety.

Data Dependency

Most types of AI systems are heavily dependent on data, sometimes in vast amounts. The quality, volume, and diversity of AI training data directly affect the performance and reliability of these systems. Poor or biased data can lead to inaccurate or unfair outcomes.

For example, AI systems used in criminal justice, such as sentencing tools, may be trained on biased historical data, leading to discriminatory outcomes. In another example, if historical arrest data disproportionately represents certain demographic groups, the AI system may unfairly flag individuals from those groups as having a higher risk of reoffending.

Governance structures must address matters of *data quality* and *security*. Organizations should establish standards for collecting, storing, and processing data to ensure it is accurate, complete, and protected.

Privacy

Many AI systems depend heavily on data about employees, customers, and other constituents. Often, these datasets contain one or more elements of PII.

For example, organizations collect PII from customers and buy more from data brokers. They may train their AI systems with this data for various purposes, including sales forecasts, targeted marketing, product and service preferences, and demographic analysis. Some organizations may violate their own *privacy policy, notice of privacy practices (NOPP)*, and/or applicable privacy laws if they don't first examine them to determine what's permitted.

We still lack a reliable method for untraining an AI once it has ingested data. This introduces a serious challenge: after the training phase is complete, complying with data protection regulations becomes almost impossible. It underscores a critical point: Data must be anonymized and safeguarded before it is ever introduced into an AI model.

Governance structures must address matters of bias, fairness, privacy, legality, and protection of PII. It is essential to establish data collection, storage, and processing standards to avoid bias, ensure compliance with privacy regulations (for example, GDPR), include notices of privacy practices (NOPPs), and protect sensitive information.

Intellectual Property

Some public AI systems, including LLMs like ChatGPT, Copilot, Grok, and Claude, have been known to train their AI models with vast amounts of information, including copyrighted content such as newspaper articles, books, images, and art. As of mid-2025, there are no established legal precedents in the United States to determine whether training AI systems with copyrighted content is ethical or legal.

For example, Getty Images filed a lawsuit against Stability AI in 2023, alleging that Stability AI used millions of Getty Images photographs, along with captions and metadata, to train its AI image-generation tools without authorization.

Existing copyright frameworks also fail to protect artist's personal styles, techniques, and methods. By design, today's AI learns the patterns in the data it is trained on, which means if an AI is trained on a specific artist's work, it may adopt their style and not necessarily replicate their exact work. Thus, copyright law leaves artists woefully unprotected.

Governance structures must employ measures to ensure that all data used for AI systems is vetted for copyright and other intellectual property issues. Similarly, AI governance also needs to ensure that an organization's workforce is acutely aware of the business rules regarding the use of public AI models and whether any organization data may be uploaded to them.

Probabilistic vs. Deterministic Outputs

Some AI systems, especially neural networks and those based on machine learning, often produce *probabilistic* outputs, meaning their predictions or decisions are based on probabilities rather than certainties. In contrast, traditional *deterministic* systems' outputs are entirely predictable given the same inputs. AI's probabilistic nature introduces an element of uncertainty, which can be challenging when outcomes have significant consequences and must be explained.

For example, based on probabilities, a healthcare AI diagnostic tool predicts a patient's likelihood of developing any particular condition. The prediction is not definitive, and relying on it could lead to unnecessary treatments or missed diagnoses if the probabilities are misinterpreted or the AI system's accuracy is not high enough.

Governance structures must address the *uncertainty* inherent in probabilistic AI systems. Management may need to define acceptable thresholds for AI decision-making, provide strict guidelines for interpreting probabilities, and ensure transparency in decision-making. Like *risk appetite* and *risk tolerance*, organizations should also determine the level of certainty required in high-risk contexts, as defined by their AI classification model.

Principles of Responsible AI

As AI becomes more integrated into everyday life, society, and critical business sectors such as energy, healthcare, finance, and criminal justice, the principles of *responsible AI* have emerged as core principles for ethical and effective uses of AI. Responsible AI ensures that AI systems are designed, developed, and deployed to respect human rights, promote fairness, and minimize risks while benefiting individuals, organizations, and society.

The remainder of this section identifies and explains the key principles of responsible AI, including ethics, fairness, safety and reliability, privacy, security, transparency and explainability, accountability, and human-centricity. These principles are the core elements that should responsibly govern AI's acquisition, development, and use. Each principle is illustrated with examples to demonstrate its practical application.

Ethics

Ethics should be considered the primary foundation of responsible AI. Ethical considerations must guide the development, deployment, and use of AI systems to align with core human values. AI systems must be designed to operate within ethical boundaries that reflect societal norms and sound moral principles.

Challenges with Ethics

AI is disrupting ethical and social norms. AI algorithms can make decisions that conflict with moral values, such as prioritizing profit over human well-being or disregarding human rights.

Like statistics, ethics can be twisted and turned to suit any outcome. Organizations must ensure that their ethical standards rest on solid foundations and are not driven by non-human-centric agendas.

For example, military organizations employ AI-powered drones that make decisions on the battlefield, with ethical implications that raise serious concerns about accountability and human dignity.

Applications of Ethics

Ethical AI requires sound, comprehensive guidelines to ensure that AI systems promote human well-being and not harm or exploit the vulnerable. AI systems should align with sound ethical principles, including human rights, dignity, and social good. Organizations should commission ethical audits, performed by objective, external parties without a stake in their outcomes. Impact assessments and stakeholder engagement are essential to identify and address ethical risks throughout the AI development lifecycle.

Questions for Further Consideration

Can an AI system that is developed without ethical considerations be considered responsible? Why or why not?

Who, inside or outside the organization, should establish ethical norms? How can they be determined to be sound and defendable?

How do ethical norms vary among societies? Can an AI system with a worldwide audience satisfy ethical norms in all locations?

Fairness

Humans seem to have a built-in instinct for fairness, as demonstrated by toddlers who are quick to point out differences in how they are treated with the exclamation, "That's not fair!" Fairness is a core component of our social fabric and is essential to responsible AI.

Fairness is a core principle of responsible AI that compels us to design and use AI systems that do not discriminate against individuals or groups based on characteristics such as race, gender, religious and political affiliation, ethnicity, socioeconomic status, and disability. When achieved, AI systems minimize or eliminate *bias* and promote equitable outcomes for all.

Challenges with Fairness

AI systems can unintentionally perpetuate or even amplify existing biases, particularly if their training data reflects historical inequities. Intentional or not, biases can lead to discriminatory outcomes, particularly when AI is used in decision-making processes, including hiring, lending, or criminal justice.

For example, a large retail organization employed an AI-powered system for screening new employment applications. After a short time, management discovered that the screening system rejected female applicants far more than male applicants. An investigation into the matter found that the training data used in the system consisted of the resumes and profiles of existing employees, who were predominantly and disproportionately male.

Not all societies and cultures have the same norms regarding fairness. Multinational organizations may need to understand potential differences and design their AI processes and systems accordingly.

Applications of Fairness

To ensure that AI systems exhibit fairness, organizations must ensure that the data used to train AI systems is diverse and representative, regularly audit AI systems for bias, and apply techniques such as *bias mitigation* to prevent further discrimination. AI systems should ensure *equitable access* and opportunities for all persons, regardless of their demographic characteristics.

Safety and Reliability

Safety and *reliability* are critical principles that ensure AI systems are safe for individuals, groups, organizations, and society. AI systems should function as intended, even when facing unexpected inputs in unpredictable environments. AI systems must also be robust to handle situations associated with errors or malfunctions.

Challenges with Safety

AI systems, especially autonomous systems with lives at stake, can pose significant risks if they malfunction or make bad decisions. The stakes are high in safety-critical areas like healthcare, emergency response, and autonomous transportation.

In autonomous vehicles, AI makes real-time navigation and driving decisions. A malfunction or failure in the AI system, such as the inability to identify a pedestrian in poor lighting conditions, could lead to accidents, injuries, and even fatalities.

Applications of Safety and Reliability

AI systems related to life safety should be developed with strict and comprehensive requirements. Those AI systems should be subject to rigorous testing and validation under various conditions. Further, such AI systems should be designed with *fail-safes,*

interruptability, and *fault tolerance*. Organizations should employ *continuous monitoring* of AI systems in life safety contexts to identify and address emerging risks.

Privacy

Many AI systems process large amounts of sensitive personal data and PII. To establish and maintain trust, organizations must take extra precautions so that all uses of AI comply with applicable privacy regulations. To ensure the *privacy* of PII, organizations must carefully track all uses of PII and provide means for responding to inquiries, corrections, and removal requests as required.

Challenges with Privacy

Even before AI, organizations were challenged to constrain themselves from using employee and customer data for uses beyond what was stipulated in privacy policies and applicable regulations. Some organizations behave as though there are no rules regarding AI systems and the data used to train them. Without strong data governance and AI governance functions, misbehavior may continue unabated until these practices can be reined in.

For example, an organization's mobile health app that uses AI to track fitness goals may inadvertently expose users' sensitive health data if its design is not analyzed for compliance with policies and regulations. Such violations could lead to public sanctions and brand damage.

Applications of Privacy

AI systems that use PII must be designed to comply with all applicable privacy laws. Privacy techniques, including *data minimization* (that is, collecting only the data necessary for the task at hand) and *deidentification* (various techniques to scrub datasets), should be employed. Security measures, discussed next, should be a part of an AI system's design.

Security

Many AI systems process large amounts of sensitive data, including personal information, intellectual property, and copyrighted content. Protecting this data from unauthorized access and compromise is a cornerstone of trust and is essential for applicable privacy laws (which require both the proper handling and use of personal data and its protection from theft).

Protecting personal information, intellectual property, and other content is foundational through long-established cybersecurity practices. This book summarizes these techniques, which are explored in great detail in other books published by the author.

For example, a retail organization is pushing to implement AI to improve the efficiency of its customer service function. Like many of its peers, the organization already struggles to understand where all its critical data lies and how it is used. This problem is only made worse when some of this data is also used to train its AI systems, which adds to the proliferation of sensitive data it struggles to protect.

Challenges with Security

Many organizations struggle to protect sensitive information and critical systems, as cybersecurity is often prioritized lower than profitability, time-to-market, and other business survival imperatives. Organizations lacking effective IT and cybersecurity governance will likewise struggle with AI if these foundational capabilities are substandard. Yet, organizations face tremendous pressure to innovate and drive efficiency using AI, as their peers are likely already doing. This is exacerbated by the fact that modern AI is subverting traditional security and trust boundaries.

Applications of Security

Organizations should first inventory their information systems and sensitive datasets, then perform risk assessments to identify risks of potential compromise and loss. Then, with management support, additional safeguards can be applied or improved to bring cybersecurity to a level of due care. This process is a standard, although brief, blueprint for modern cybersecurity.

Transparency and Explainability

The term *transparency* broadly refers to the ability of an organization to understand how its AI systems function. More specifically, *interpretability* is an organization's ability to understand how it reaches its conclusions. *Explainability* is the characteristic of an AI system that allows an organization to provide understandable justifications for the actions the system makes. These critical principles help build trust and ensure accountability, particularly for AI systems making impactful decisions.

Challenges with Transparency and Explainability

Many AI models, including neural networks and those employing deep learning, operate as "black boxes," making it difficult for an organization to understand its outputs and decisions. This opacity leads to trust issues and challenges to AI systems' decisions.

For example, a community bank has recently implemented an AI tool to examine loan applications. When the AI system rejects a loan application, the organization may not provide a concise reason for the rejection. The applicant is dissatisfied with the lack of a credible explanation and may complain to regulators.

Applications of Transparency and Explainability

To avoid situations in which an organization cannot adequately explain the reasoning behind an AI system's decisions and outputs, organizations must prioritize the development of *explainable AI (XAI)*, where AI models provide understandable reasons for their outputs. While this may appear to hamper an organization wishing to provide the most powerful AI capabilities, using XAI in decision-making contexts can help organizations avoid situations that can lead to financial, social, operational, and reputational damage.

Accountability

Accountability is that organizational trait that ensures that individuals and organizations are held responsible for the outcomes of their AI systems. The roles and responsibilities for AI system oversight and recourse are essential when problems arise.

Challenges with Accountability

AI systems often operate autonomously and make decisions independently, but this does not absolve anyone of responsibility when something goes wrong.

For example, a healthcare organization's AI-enhanced diagnostic system makes an incorrect diagnosis, leading to patient harm. The healthcare organization tried to blame the diagnostic system and its manufacturer, but the outcome in arbitration results in accountability resting solely with the healthcare provider, because it selected and configured the device, and it failed to conduct additional follow-up to confirm the diagnosis.

Applications of Accountability

Organizational *policy*, culture, and AI governance frameworks must establish and enforce clear accountability, including assigning responsibility for every aspect of AI system design, deployment, and monitoring. In some cases, organizations should establish mechanisms for *recourse* (including dispute resolution) when harm is alleged. Organizations also need to be transparent about the workings of their AI systems and provide individuals with avenues to challenge AI system decisions.

It's easy to pin the blame on AI developers and hold them accountable for their AI systems' actions. However, if developers are solely accountable for the costs of security, their products may become less attractive, which leads organizations to opt for cheaper, less secure alternatives. That's why accountability should ultimately lie with the deployers. When organizations are held responsible for the behavior of the AI systems they deploy and use, security quickly becomes a much higher priority.

The Culture of Accountability

Organizations are all over the place on the topic of accountability. In other words, are staff members, managers, leaders, and executives accountable for their actions and decisions? Are they required to face consequences when things don't go as intended? Organizations that struggle are more likely to blame others (vendors, service providers, and AI systems) when problems occur. If your AI system makes a decision that gets your organization in hot water, don't blame the AI system; after all, you selected it and are accountable for all outcomes.

Human-centricity

The principle of *human-centricity* states that AI systems should be designed with human well-being and dignity in mind. AI should augment human capabilities, improve people's lives, and preserve their dignity, rather than replace or diminish human roles. AI systems should be designed and implemented in alignment with human values and priorities.

Challenges with Human-centricity

As organizations strive to improve efficiency in all forms, they may cut too close to the bone and prioritize financial efficiency over human values. Organizations employing AI systems focusing on profit margins might forsake the well-being of their workers or customers.

For example, a factory has implemented AI to improve production efficiency, which results in worker displacement. The organization may improve its financial position but at the cost of losing its community-centric reputation.

Applications of Human-centricity

Organizations must emphasize human-centered design principles, ensuring that AI systems are developed to serve human interests. This may require *impact assessments* to determine how AI systems enhance, rather than displace, human capabilities. Processes in which AI systems make decisions should include a *human in the loop*, particularly in financial services, healthcare, and justice.

Summary

Artificial intelligence simulates human intelligence with capabilities like learning, reasoning, problem-solving, perception, and language understanding.

AI systems are classified in various ways, including their scope (narrow/weak AI, general AI, artificial general intelligence—AGI, and artificial superintelligence—ASI). AGI and ASI models have not yet been developed and announced to the general public.

AI functions include reactive machines, limited memory AI, theory of mind AI, self-aware AI, and robotics. Each type of function is suited to a range of tasks.

Some of the techniques related to AI systems include machine learning, deep learning, natural language processing, and expert systems. AI system training techniques include supervised, unsupervised, semi-supervised, and reinforcement learning.

AI systems pose certain risks to individuals, groups, organizations, and societies without proper safeguards. The risks to individuals include privacy violations, bias and discrimination, cyberattacks, and loss of autonomy and control. The risks to groups include the amplification of social inequality, erosion of social cohesion, and cyberattacks. Risks to organizations include operational disruption, legal and regulatory issues, financial risk, including cost overruns and legal costs, cyberattacks, and reputation damage.

The risks to society include job displacement, economic inequality, erosion of human agency, cyberattacks, and manipulation of public opinion.

Organizations often embrace and experiment with emerging technologies to gain first-mover advantage. Historic examples include the emergence of the personal computer (PC), the Internet, and cloud computing. There were winners and losers in every generation, partly because of good planning and partially due to good luck.

For most organizations, AI systems are the most complex and impactful systems they'll ever use. This complexity and power alone should convince any organization to put governance structures in place to ensure intended outcomes, as the consequences of mishaps can be substantial.

The characteristics of AI systems, particularly complexity and opacity, challenge organizations that employ AI systems in decision-making.

AI systems can deliver speed and efficiency but may warrant a human in the loop in high-risk, consequential roles, such as medical diagnosis, employment applicant analysis, and loan application analysis.

Autonomous AI systems, such as autonomous vehicles, need safeguards to minimize the risk of impactful, unavoidable circumstances.

Organizations whose AI systems process PII must be sure to comply with all applicable privacy laws, as well as organizations' own notices of privacy practices. Further, organizations must protect the data used by AI systems to resist cyberattacks.

Responsible AI is a cornerstone principle of AI. Great power calls for great responsibility, but AI systems also bring the risk of highly consequential errors and the lure of misuse. This and several characteristics of AI call for establishing effective governance structures to ensure only intended outcomes.

Ethics is a core tenet of responsible AI, but organizations must approach this with caution, as ethics can be subjective and twisted at will.

Fairness is a key part of responsible AI, as it's vital that AI systems in decision-making roles always produce outcomes that are seen as fair to all concerned.

The safety and reliability of AI systems is yet another facet of responsible AI, particularly in contexts like autonomous vehicles and robotic surgery.

Privacy and security are a part of responsible AI, where organizations are duty-bound to comply with privacy regulations, principles, and policies, and to protect all forms of data in an AI system from cyberattacks and other forms of abuse.

Responsible AI also includes transparency and explainability when AI systems make decisions that impact others, such as employment and loan applications. Organizations must understand (and explain to affected parties) how an AI system produces its outputs. Some types of AI systems are more explainable than others; organizations must choose wisely.

Responsible AI includes accountability—organizations cannot blame AI systems or their makers when things don't go as planned. Organizations need to have a culture of accountability, so that decision-makers take their roles seriously, since they will be bound by their consequences.

Finally, responsible AI includes human-centricity. AI can vastly improve the efficiency of its business processes, but when it comes to interaction with employees and customers,

it's unwise to make them "feel like a number." Further, AI systems in decision-making roles should involve a human in the loop to ensure fairness to affected parties.

Exam Essentials

Be able to explain the various types of AI systems. AI systems are classified and categorized in various ways, based on their capabilities, functions, and how they are trained and used. The types of AI systems include weak/narrow AI, general AI, and artificial superintelligence (not all exist today). AI functions include reactive machines, limited-memory AI, theory of mind AI, and robotics.

Know the characteristics of AI that necessitate governance. While AI systems can improve organizational efficiencies and quality, they are so potentially powerful that the consequences of unintended outcomes can destroy an organization. These characteristics include complexity, opacity, autonomy, speed and scale, and potential misuse and harm.

Know the principles of responsible AI. While AI systems can improve organizational efficiencies and quality, they are so potentially powerful that the consequences of unintended outcomes can harm citizens, groups, organizations, and society as a whole. Thus, AI systems must be responsibly developed. The principles of responsible AI include ethics, fairness, privacy, security, transparency, explainability, and accountability.

Review Questions

The following questions are designed to test your understanding of this chapter's material. For more information on how to obtain additional questions, see this book's Introduction. You can find the answers in Appendix A.

1. The Netflix Recommendation Engine and Google Search Autocomplete are examples of:
   - **A.** AI agent
   - **B.** General AI
   - **C.** Narrow AI
   - **D.** Training techniques
2. An organization has developed an AI system that receives and analyzes input from sensors and takes action as needed. This system is an example of:
   - **A.** AI agent
   - **B.** Robot
   - **C.** General AI
   - **D.** Operational technology
3. An EU-based retail organization retains data about individual transactions and maintains databases with detailed customer information. The organization wants to build an AI system to analyze its customers and transactions, to help better understand sales history, and to develop better sales and marketing strategies. What interaction, if any, is required with customers before launching this project?
   - **A.** Customers can opt out of this activity.
   - **B.** Customers must be notified of this activity.
   - **C.** No customer interaction is required.
   - **D.** Customers must explicitly consent to this activity.
4. An IT department is building an AI system and will begin training it with labeled data. What kind of training is this?
   - **A.** Supervised learning
   - **B.** Unsupervised learning
   - **C.** Semi-supervised learning
   - **D.** Reinforcement learning
5. A regional bank is building an AI system to help analyze loan applications and make approval and interest rate recommendations. What kind of harm must the bank ensure is not realized?
   - **A.** Financial
   - **B.** Social cohesion
   - **C.** Privacy
   - **D.** Bias

6. A regional bank is building an AI system to help analyze loan applications and make approval and interest rate recommendations. What safeguard should the bank implement to ensure fairness?
   - **A.** Event detection
   - **B.** Supervised learning
   - **C.** Human in the loop
   - **D.** Hallucination detection

7. A social media platform uses an AI algorithm to select items to appear in users' data feeds. What potential harm could result if the AI algorithm is biased?
   - **A.** Job displacement
   - **B.** Manipulation of public opinion
   - **C.** Erosion of human autonomy
   - **D.** Privacy

8. An organization has implemented a neural network AI system to review job applications and make interview recommendations. An employment candidate recently applied for a position, was rejected by the AI system, and requests an explanation for the rejection. What potential challenge is the organization about to face?
   - **A.** Security incident
   - **B.** Privacy violation
   - **C.** Difficulty in providing a reason for the rejection
   - **D.** Economic inequality

9. The highest priority in the design of the AI system controlling an autonomous vehicle should be:
   - **A.** Privacy
   - **B.** Safety
   - **C.** Fairness
   - **D.** Security

10. The highest priority in the design of an AI system visually identifying returning customers in a retail store is:
    - **A.** Security
    - **B.** Fairness
    - **C.** Safety
    - **D.** Privacy