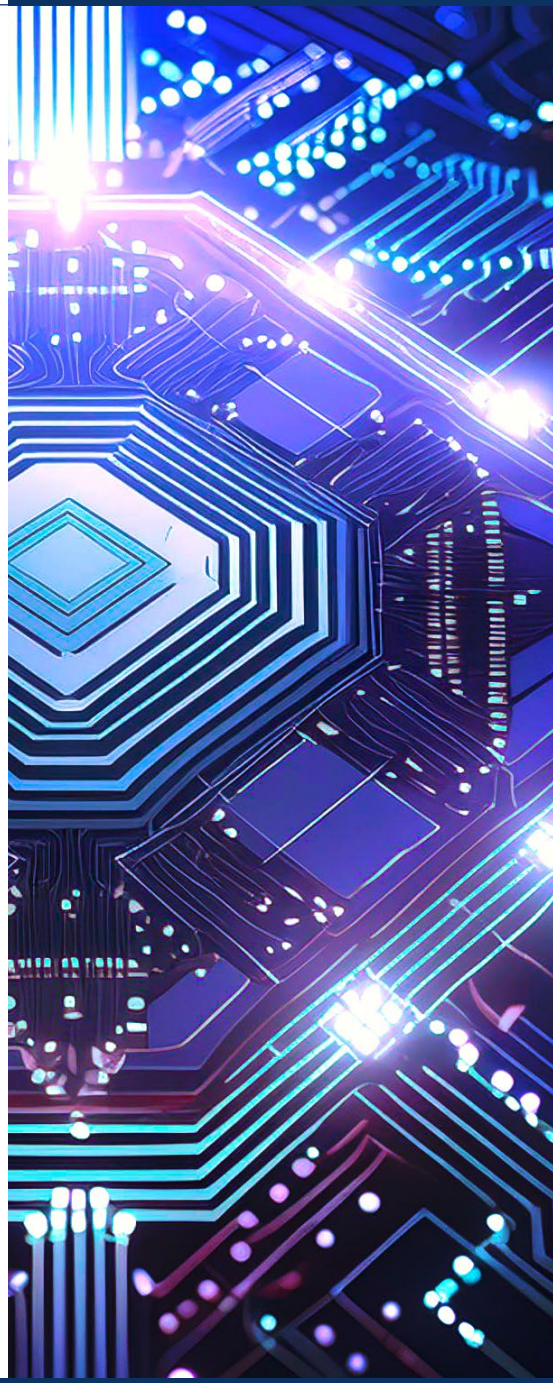
A JOINT PUBLICATION BY

Australian Institute of Company Directors

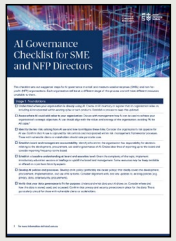
UTS Human Technology Institute

A Director's Introduction to AI

Contents



How to use this guide

Having considered all the boards on which you serve, select what applies to you:	What we suggest you read
☐ I know about ChatGPT, but I don't know any other types of AI ☐ I am not clear how AI is different to other technologies ☐ I am unsure about the key legal obligations applying to AI use ☐ I am not clear about the key risks or opportunities arising from AI ☐ I do not know the underlying principles of safe and responsible AI	 A Director's Introduction to AI
☐ I understand the difference between General AI and Narrow AI ☐ I understand how AI is different to other technologies, but am unclear how this impacts governance ☐ I am unsure about where AI is used within my organisation ☐ I am unsure about what questions to ask management about the governance and use of AI and how to assess the quality of management's responses	 A Director's Guide to AI Governance
☐ I am a director of a SME or NFP and do not know how to implement AI governance	 AI Governance Checklist for SME and NFP Directors

Resource purpose, audience & structure

This resource is intended to introduce directors to key AI concepts, and is structured into three chapters:

- **Chapter 1** provides directors with an introduction to what AI is, how it is being used and its relevance for directors.
- **Chapter 2** outlines the key opportunities and risks of AI use.
- **Chapter 3** examines current regulatory obligations related to AI systems and the shifting regulatory environment locally and internationally.

It is not intended to 'cover the field', but to develop a foundational understanding of AI governance.

The resource lays the foundation for directors to apply AI governance principles. This is set out in Part 2 of this series, 'A Director's Guide to AI Governance.'

Executive summary

- There are two main types of AI systems: (1) General AI systems (which include Generative AI such as ChatGPT) and (2) Narrow AI systems.
- General AI and Narrow AI systems are subject to different risks and present different governance challenges. Both types of AI require additional consideration and oversight from management and directors
- Managing AI systems can be particularly challenging because of their sophisticated pattern recognition capabilities, which operate at a large scale and pull from vast datasets to generate complex outputs.
- AI use within an organisation may not be obvious, which compounds the governance challenge.
- Increasingly, AI is being deployed in core organisational functions such as strategy, corporate finance and risk. This trend is likely to continue, increasing the need for boards to implement safe and responsible AI governance.
- Directors are ultimately responsible for the oversight of risk management throughout the organisation. This includes risk arising from AI.
- Existing legal obligations in the areas of privacy, consumer protection, intellectual property, cyber security, anti-discrimination, duty of care and work, health and safety continue to apply, and will be relevant to AI use.
- The regulatory landscape is evolving rapidly. While regulatory approaches differ, a common set of safe and responsible AI principles underpin reforms locally and internationally.

CHAPTER 1: AI and the relevance for directors

KEY POINTS:

- There are two main types of AI systems: General AI (or General Purpose AI) and Narrow AI systems. They are subject to different risks and present different governance challenges.
- Managing AI systems can be particularly challenging because of their sophisticated pattern recognition capabilities, which operate at a large scale and pull from vast datasets to generate complex outputs. This can make their decisions difficult to explain.
- AI use within an organisation may not be obvious, which compounds the governance challenge.
- Increasingly, AI is being deployed in core organisational functions such as strategy, corporate finance and risk. This trend is likely to continue, increasing the need for boards to implement safe and responsible AI governance.

1.1 WHAT IS AI?

The definition of AI adopted by the International Organisation for Standardization and the International Electrotechnical Commission ISO/IEC 22989 is:

> *An engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives.*

1.1.1 How is AI different from other technology?

AI is a special form of digital software that is particularly good at predicting outputs, optimising, classifying, inferring missing data, and generating new data.

AI systems can often outperform traditional software, and as a result offer significant productivity, efficiency and customer experience benefits.

AI is also more versatile and scalable than traditional software because it can be replicated and adapted to new contexts at a relatively low cost. As a result of these advantages, AI is increasingly being deployed across organisational teams and functions.

However, the differences between traditional software systems and AI systems also impacts governance approaches.

Traditional software systems are built from explicit rules coded by developers, such that their behaviour is inherently more predictable and understandable (even if the software itself is complex).

By contrast, AI systems are often created by defining an objective and using historical data to create an AI model that may rely on billions of inferred connections between data points to achieve its objective. This process means that **it can be extremely challenging to replicate, explain or test an AI system's output.**

BOX 1: The role of data in AI systems

Data is the foundation of AI systems. Data, including personal information, is collected and used to train AI systems. It is both an input and an output of a deployed AI system.

The selection of data, particularly its quality, quantity, and representativeness, will significantly affect the performance of AI systems.

Through the ongoing collection of data and feedback loops, the accuracy and efficiency of AI systems should improve over time.

BOX 2: What kinds of systems are usefully defined as AI?

- **Machine learning:** a broad set of models that have been trained on pre-existing data to produce useful outputs on new data.
- **Expert systems:** systems that use a knowledge base, inference engine and logic to mimic how humans make decisions.
- **Natural language systems:** models that can understand and use natural language and speech for tasks such as summarisation, translation, or content moderation.
- **Facial recognition technologies:** systems that verify a person, identify someone, or analyse personal characteristics using facial data drawn from photos or video.
- **Recommender systems:** systems that suggest products, services or information to a user based on user preferences, characteristics, or behaviour.
- **Automated decision-making systems:** systems that use data to classify, analyse and make decisions that affect people with little or no human intervention.
- **Robotic process automation:** systems that imitate human actions to automate routine tasks through existing digital interfaces.
- **Virtual agents and chatbots:** digital systems that engage with customers or employees via text or speech.
- **Generative AI:** systems that produce code, text, music, or images based on text or other inputs.
- **AI-powered robotics:** physical systems that use computer vision and machine learning models to move and execute tasks in dynamic environments.

1.1.2 Different types of AI

Box 2 provides a non-exhaustive list of systems that meet the definition of AI above.

General AI (or General Purpose AI) and **Narrow AI** are two sub-categories of AI (see Table 1).

TABLE 1: Key differences between General AI and Narrow AI

Type of AI system	Description[1]	Examples
General AI (or General Purpose AI)	An AI system that can be used for a broad range of tasks, both intended and unintended by developers. This includes Generative AI.	Text generation (i.e. GPT-4, Gemini), image generation (i.e. DALL.E, Midjourney), programming code generation (i.e. Codex).
Narrow AI	An AI system trained to deliver outputs for specialised, constrained tasks and uses to address a specific problem.	Search engines (i.e. Google, Bing), facial recognition (i.e. Apple Face ID), recommender systems (i.e. Amazon, Spotify, Netflix).

1 ISO, 2022. ISO-IEC-22989 Artificial intelligence concepts and terminology.

As discussed further in Chapter 2, General AI (including Generative AI) and Narrow AI present slightly different governance challenges (see Box 9).

1.1.3 How do I know when AI is being used in my organisation?

AI use is not always obvious. This makes its use more difficult to govern. **Box 3** sets out terms that may indicate that AI systems are in use within an organisation.

As AI advances rapidly, corporate leaders would be well-served to take a broad view of what constitutes an AI system within their organisation.

BOX 3: Key terms to listen out for to identify potential AI use within your organisation

The use of AI by organisations is not always clear to executives and directors. In addition to the kinds of systems listed in **Box 2**, common terms to listen for which ***may*** indicate the use of AI and warrant further investigation include:

- **Model or algorithm** (e.g. a specialised piece of software designed to provide a recommendation, optimise a system, or prioritise an action).
- **Training data** (e.g. data used to train or fine-tune an AI algorithm).
- **Data analytics** (e.g. a set of data transformations to classify consumer profiles).
- **Predictive analytics** (e.g. using data to predict future trends or events).
- **Prescriptive analytics** (e.g. analysing data to identify the optimal course of action).
- **Process automation** (e.g. the use of robotic process automation to perform repetitive tasks).
- **Automated decision-making** (e.g. the use of a set of rules or a self-learning algorithm to make a decision, such as providing a risk classification or approving a further action).

1.2 HOW AND WHY IS AI BEING USED BY ORGANISATIONS?

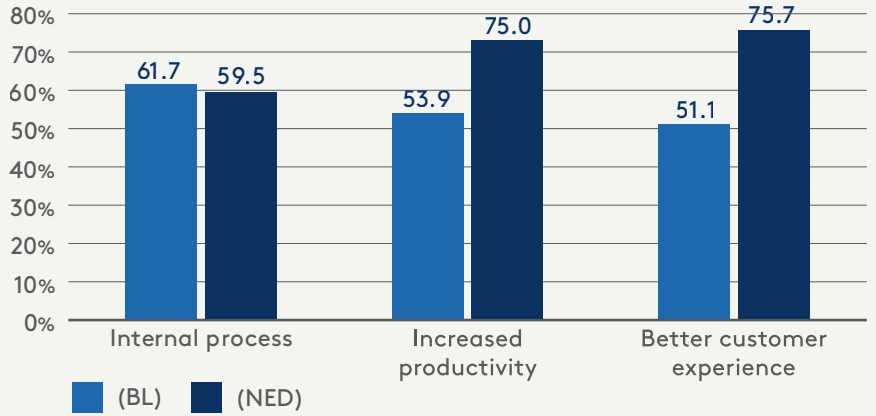
AI is rapidly becoming an essential part of how organisations operate. Research from Human Technology Institute (HTI) conducted with business leaders and directors in 2023 found that almost **two-thirds** of Australian organisations are already using, or actively planning to use AI systems to support a wide variety of functions.

Organisations are introducing AI systems to secure a range of benefits, including:

- reducing costs;
- enhancing productivity;
- improving customer experience; and
- delivering new business growth.

HTI's data indicates that **non-executive directors** tend to place more **focus on the opportunity for AI systems to serve customers better**, while managers tend to see greater value in deploying AI systems to achieve process efficiencies (**Figure 1**).

FIGURE 1: Top expected benefits of AI use by Business Leaders (BL) and Non-executive Directors (NED) surveyed by HTI in 2023



While Narrow AI systems have traditionally been the domain of data analytics teams, cyber security systems or other back-office functions, the rising capabilities and flexibility of AI systems mean they are increasingly being used in ways that touch stakeholders directly.

AI systems are undergoing significant changes in their application. **Three of the top five** priority areas for AI system use directly impact consumers or employees, including customer service, marketing and sales, and human resources.[2]

AI is becoming increasingly used in three key areas:

1

IMPROVING THE CUSTOMER AND EMPLOYEE EXPERIENCE

AI is being used to extend and augment the reach and output of employees in a way that can improve the customer experience and reduce the drudgery of mundane tasks (thereby freeing up employees for higher value-add work). We detail some of the benefits of AI use in **Chapter 2** (**section 2.1**). For example, Telstra is using Generative AI systems to support frontline teams and answer complex customer queries, while an AI-driven dashboard in NSW hospital emergency rooms helps doctors identify patients at a high risk of sepsis.

2

INCORPORATION INTO NEW PRODUCTS AND SERVICES THROUGHOUT THE VALUE CHAIN

AI is being bundled into products and services that organisations procure through technology partners. This means it is often used by employees and across supply chains in ways that are often not fully visible. For example, a February 2024 survey of 1,000 office workers commissioned by Salesforce found that 53 per cent of Australian professionals are actively using or experimenting with Generative AI at work. Not all of this employee use of AI is disclosed (known as 'Shadow IT' or 'shadow AI use' (see **Box 3** in **Section 1.3** of A Director's Guide to AI Governance)), which creates risks and governance challenges.

3

INCORPORATION INTO CORE BUSINESS FUNCTIONS

AI systems are being applied closer to the 'core' of organisations, with some of the most rapid growth in strategy, corporate finance, and risk functions. For example, a 2024 NVIDIA survey found that risk management was the second-highest current use and top investment domain for AI systems in the financial services sector.

2 Lauren Solomon and Nicholas Davis, The State of AI Governance in Australia (HTI Report, 2023).

CHAPTER 2: AI opportunities and risks

KEY POINTS:

- AI use can produce benefits and opportunities as well as risks and harms.
- Key benefits of using AI systems include increased productivity, quality improvement, new products and services, and an improved customer and employee experience.
- Many of the potential harms to consumers or employees from AI system misuse or failure are foreseeable and capable of mitigation.
- Without appropriate controls, AI systems tend to negatively and disproportionately affect vulnerable and marginalised populations. Organisations need to ensure that processes are in place to identify and prevent these harms.

2.1 AI OPPORTUNITIES

AI systems promise a range of significant benefits for organisations. These include:

INCREASED EFFICIENCY AND PRODUCTIVITY

Some AI systems can reduce the time burden of administrative tasks through new forms of automation. Others allow employees to expand their output and add additional value, helping teams to analyse trends, summarise existing content, and generate new content. A 2023 experiment designed by MIT Sloan found that when Generative AI was particularly suited to a task, it could enhance worker productivity by approximately 40 per cent.

REDUCTION IN ERROR AND QUALITY IMPROVEMENTS

While AI systems are extremely prone to errors when input data or queries fall outside their core competency, for well-known mechanical or repetitive tasks, particularly those involving pattern recognition, they can perform significantly better than other approaches, including human experts.

NEW PRODUCTS AND SERVICES

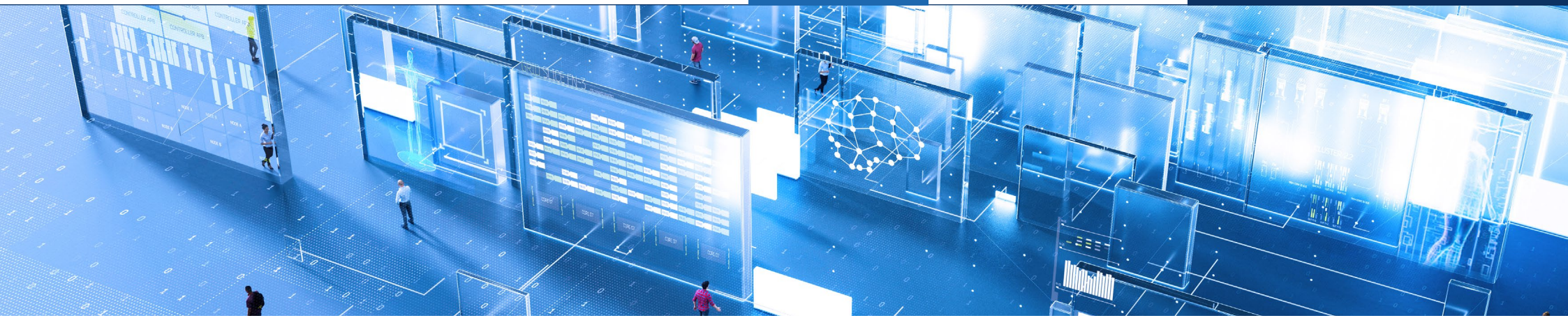
Both Narrow and General AI systems can support organisations with a range of innovation-related tasks, including helping organisations identify, predict demand, design, prototype, and test new products and services.

IMPROVED CUSTOMER EXPERIENCE

Thanks to their ability to engage in natural language and scale digitally, General AI systems are helping to reduce customer wait times, improve accessibility of existing information, and personalise the customer experience.

IMPROVED EMPLOYEE EXPERIENCE

AI can reduce the time and cost spent by employees on administrative tasks to allow focus on value-add work and innovation. Some AI (such as Generative AI) can also guide workers through more complex tasks and can assist in problem-solving.



Capitalising on the opportunities above requires an investment of time and resources. Successful implementation of AI also relies on high levels of trust and engagement from customers and employees, supporting infrastructure (including effective data governance), and users with the necessary skills and training.

Conversely, **inaction or failing to seize the opportunities offered by AI can present significant risks for organisations** (see section 2.2.3). Early adopters of AI systems have gained, and are continuing to gain, competitive advantages.[3] For example, AI-driven search, pioneered by Google, signficantly impacted advertising strategies, and AI systems were central to the disruption of the taxi industry by rideshare companies.

Sectoral differences are also emerging in AI adoption and use. In Australia, research indicates that aerospace, defence and security, mining, energy and resources, agriculture, health, and transport were the top industries serviced by AI firms in 2021.[4] Globally, the industries leading the adoption of AI technology in 2023 were technology, financial services, health, transport and education.[5]

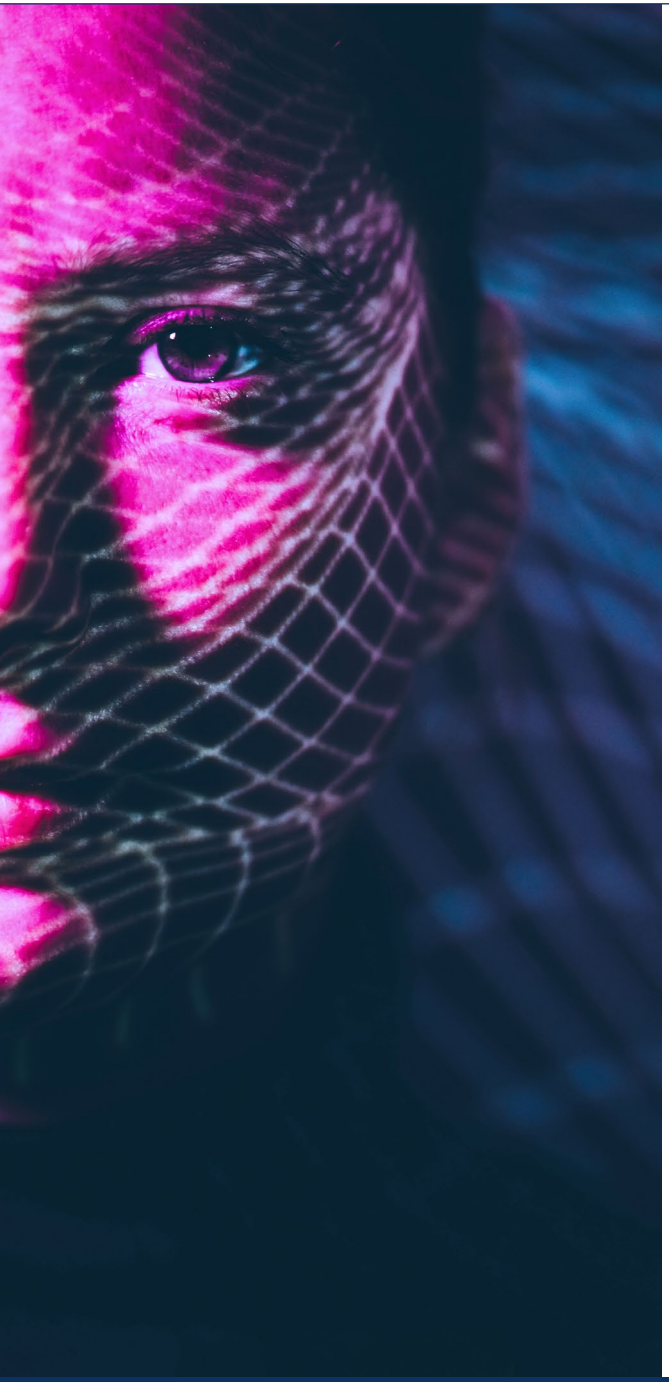
BOX 4: Potential economic impact of Generative AI on the Australian economy

In July 2023, Microsoft and the Tech Council of Australia issued a report on the economic impact of Generative AI on Australia.

The report found that Generative AI could add between $45 to $115 billion in GDP to the Australian economy annually.

The majority (70 per cent) of these gains are estimated to come from productivity improvements - it is estimated that Generative AI has the potential to automate or augment 44 per cent of an average worker's tasks. The remainder comes from new jobs and new products, services and businesses, with the biggest opportunities in healthcare, manufacturing, retail and professional and financial services.

3 McKinsey, The state of AI in 2022 — and a half decade in review (Report, December 2022).
4 Austrade, The 2021 Australian Artificial Intelligence Export Survey (Report, 2021).
5 John Mangan, Australia's AI Imperative: The economic impact of artificial intelligence and what's needed to further its growth (Kingston AI Group Report, 2024).



2.2 AI HARMS

The specific characteristics of AI systems that set them apart from traditional software also mean that they can amplify existing harms while creating new ones that may affect individuals, organisations and/or society. Table 2 summarises these potential harms.

TABLE 2: Potential harms to individuals, organisations and society from AI systems

Harm to individuals (consumers, employees, members of the public)	• Physical, psychological, economic, or reputational harm • Misleading advice or information • Violation of civil liberties • Breach of privacy • Unlawful discrimination and exclusion • Unfair treatment
Harm to organisations	• Commercial losses • Reputational damage • Regulatory sanctions
Harm to society	• Job displacement • Economic inequality • Large-scale damage to public health, infrastructure or essential services • Environmental damage • Social and political manipulation • Discrimination and oppression of minority groups

2.2.1 Harm to individuals and the importance of vulnerable communities

As **Chapter 3** details, there are a wealth of existing laws relevant to an organisation's use of AI. Some of these may result in liability for organisations – and directors personally – if individuals are harmed as a result of AI systems.

Research indicates that AI harms tend to disproportionately affect vulnerable and marginalised communities. The reasons for this are complex, but are often driven by systemic biases that exist in the data used to train AI models, the design, engineering and modelling processes, and the contexts in which AI models are ultimately deployed by decision makers.[6]

The under representation of women or people of colour in data can lead to decreased accuracy of AI systems, such as facial recognition technologies or computer-aided diagnosis systems, including medical image interpretation.

Directors should be mindful of the impact of AI use on vulnerable and marginalised individuals, and consider ways to mitigate this (see practical steps in A Director's Guide to AI Governance).

BOX 5: The Risk of Bias

Bias is one of the most well-documented concerns related to AI systems. This issue arises because of the potential for AI systems to inherit and amplify biases present in real-world data.

AI systems are also prone to bias due to their reliance on historical data for training. Bias may emerge from:

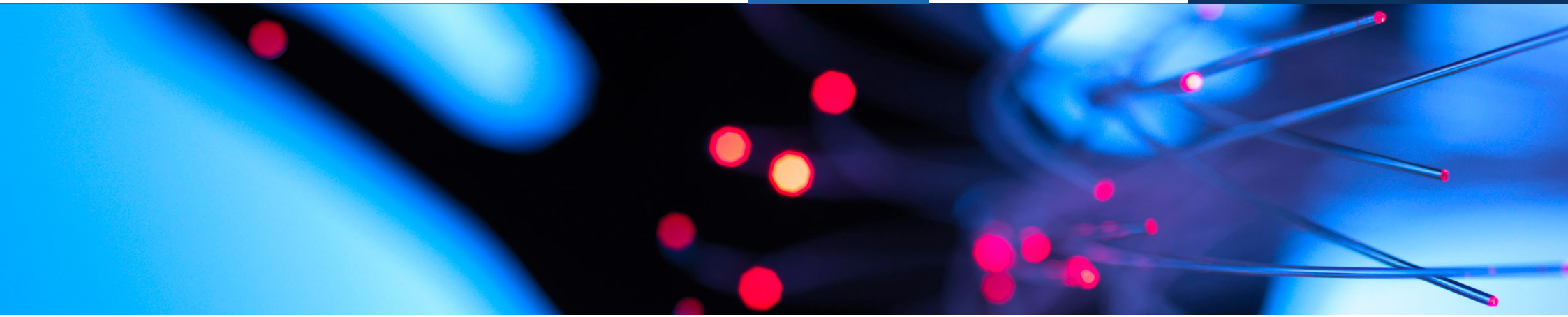
- pre-existing biases present in the real world;
- the use of non-representative data sets; and
- the selection of algorithm approaches or objective functions that intrinsically embed the bias of the development team.

Directors should be vigilant that the use of an AI system may create outcomes that are unlawful and discriminatory and disadvantage individuals or groups based on protected characteristics such as their age, race, sex, or disability.

Addressing bias in AI systems is not straightforward. Technical solutions alone are often insufficient to fully rectify the biases embedded within real-world data. The Australian Human Rights Commission's technical paper on addressing algorithmic bias provides a comprehensive examination of this issue, highlighting the limitations of relying solely on technological fixes to address biases in AI systems.

In light of this, directors should be aware of, and champion approaches that combine technical, operational, and governance practices to help mitigate the risk of bias and ensure AI systems are developed and deployed in a fair and human-centred manner.

6 Reva Schwartz et al, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (NIST Special Publication 1270, 2022).



2.2.2 Harm to society

Collective harms that can arise from AI system misuse or failure include social and political manipulation, new forms of technological unemployment, and the systemic oppression or exclusion of minority groups. When used at scale by those with broad reach (such as governments and essential service providers) even relatively small errors or biases in a system can cause large harms when scaled across groups.

While such society-wide effects are the purview of government policy, directors should be aware of these macro-level harms and how issues such as AI system job displacement may be viewed by stakeholders.

BOX 6: Workforce impact and job displacement

Research estimates that 300 million full-time equivalent workers are susceptible to automation as a result of AI systems.[7] However, the research finds AI systems are more likely to augment workers by automating some tasks, but not replace them outright. Further, AI is creating new jobs, bringing opportunities for the retraining and redeployment of workers.

A World Economic Forum **report** states that AI is expected to be adopted in 2023-2027 by 75 per cent of companies surveyed. Despite the workforce transformations that many anticipate being driven by AI, only 25 per cent of these organisations expect it to create net job losses. More than 50 per cent of organisations expect it to create net job growth.

As AI transforms the way people work, it also offers the possibility of improving worker satisfaction. A **survey** by Microsoft indicates that whilst 49 per cent of people are worried AI will replace their jobs, 70 per cent would happily delegate work to AI to ease their workloads.

Organisations should engage with employees about the impact of AI on their roles and the potential for AI to assist worker productivity and efficiency and to provide skills to allow them to gain opportunities for retraining.

7 Jan Hatzius et al, The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani), Goldman Sachs (online, 26 March 2023).

2.2.3 Harm to organisations

It is important to recognise that there are risks for organisations at two levels:

- risks arising from AI system investment and use; and
- risks arising from underinvestment and a lack of adoption.

AI misuse or system failures can create and amplify a range of **commercial, reputational and regulatory risks to organisations**.

FIGURE 2: Risks to organisations from AI use

On the other hand, a lack of investment in AI capabilities also leaves organisations vulnerable to a range of other risks, such as a lack of competitiveness, higher costs, lack of new product and service delivery, poorer consumer service, as well as talent acquisition and retention challenges.

The risks of action and inaction must be carefully weighed by directors alongside the organisational strategy and the risk appetite of the organisation (discussed further in A Director's Guide to AI Governance).

BOX 7: Sustainability, ESG and AI

AI systems raise risks and harms that may be captured under environmental, social and governance (ESG) or sustainability frameworks. For example, the training of Generative AI models can have an environmental impact given the significant energy and water it requires. The potential of AI to disproportionately impact vulnerable persons is also a 'social' risk (the 'S' within ESG).

Addressing ESG matters will necessarily address some of the risks and harms of AI systems.

In November 2022, AICD partnered with Herbert Smith Freehills to publish, under the Climate Governance Initiative (CGI) Australia banner, a 'Bringing together ESG' resource to help boards develop appropriate governance structures to effectively oversee sustainability issues.

However, given the specific challenges, harms and opportunities of AI systems, ESG frameworks alone are not sufficient. There must be broader consideration of AI systems for their effective governance (see A Director's Guide to AI Governance).

BOX 8: Generative AI and cyber security risks

Generative AI systems are likely to give rise to specific cyber security risks and undermine existing controls. For example, deepfake AI tools could be used to generate realistic synthetic media that impersonates individuals for the purposes of identity theft, fraud or spreading misinformation about a particular organisation. To mitigate these risks, boards should oversee the strengthening of cyber security controls, such as data encryption, access controls, employee training and regular external cyber security audits.

2.3 KEY SOURCES OF AI RISKS AND HARMS

AI risks and harms are created because of the way AI systems *perform and behave*, as well as how they might be *used*. Table 3 outlines some examples of how such risks arise.

TABLE 3: Key sources of AI risk for organisations

Key sources of AI risk	Examples
AI system failures – where systems create harm because they fail to perform as intended	• Poor system performance • Biased system performance • System fragility or unreliability • Security failures or vulnerabilities
Malicious, misleading, reckless, or inappropriate use – where systems are deliberately used (whether by the organisation or external parties) in a way which creates or amplifies risk of harm	• Misleading advice • Misinformation at scale • Unfair or extractive use • Opacity and lack of interpretability • Weaponisation • AI-powered cyber attacks • Fraudulent and unlawful use e.g. scams • Financial market manipulation • Excessive deployment • Deployment on vulnerable individuals

Risk management frameworks must identify and mitigate risks of system failures and misuse – this is discussed further in A Director's Guide to AI Governance.

Table 4 sets out five characteristics that impact the potential for an AI system to cause harm. Box 9 shows how these five characteristics create different levels of challenge across Generative AI (a subset of General AI) and Narrow AI systems.

TABLE 4: Factors that drive harms from system failure, misuse, or inappropriate use

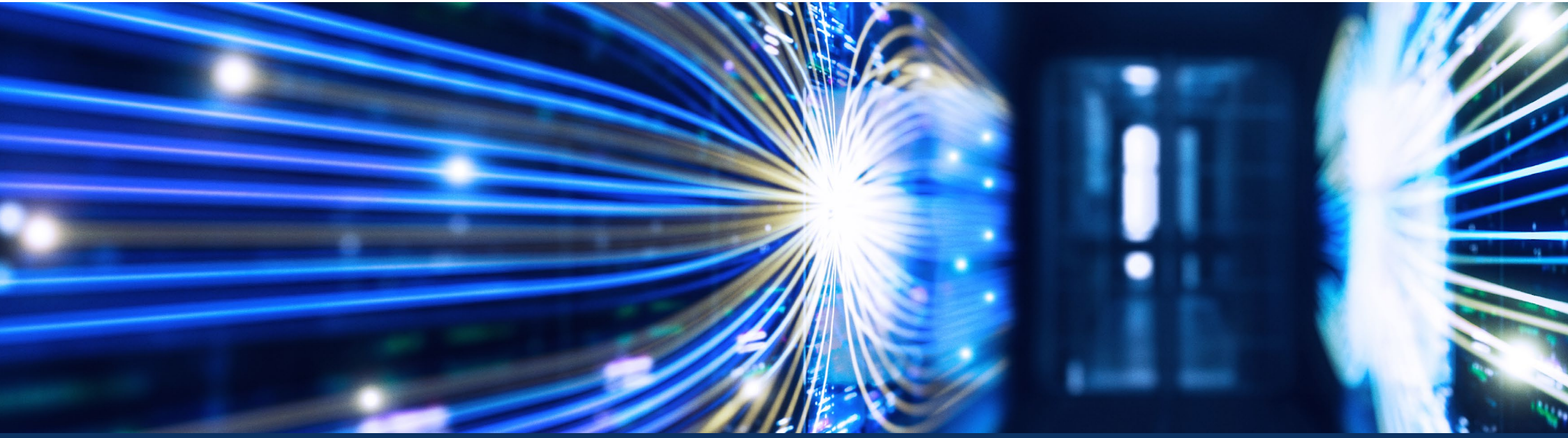
Factor	Relevance to harms
Purpose	An AI system's potential for harm is sensitive to its intended and actual purpose. If a system is used for a highly consequential purpose – for example, an output that has a legal or similarly significant effect, or a system that controls critical infrastructure – its potential to cause harm will be elevated.
Context	The potential for harm is also a function of the individuals or groups with whom the system interacts. The risk of harm is elevated if vulnerable individuals, such as children or minorities who are subject to, or require, additional forms of legal protection, are exposed to a system or are the subjects of its outputs.
Data	The risk of harm will rise when confidential, personal or sensitive information is used in an AI system's training or operation, data quality or provenance is unclear or unverifiable, if live data is ingested, and if large amounts of data are used.
Technical architecture	The choice of AI model or software, the surrounding system elements, and the quality of supporting infrastructure will also influence the risk of harm from an AI system. For example, the fact that Generative AI systems can be induced to produce content that is inappropriate in a work environment makes them inherently more risky than alternatives.
Level of automation	The potential for harm rises when an AI system is triggered automatically or produces outputs that are fed into other systems with little or no human verification, particularly when tied to a consequential purpose.

BOX 9: Emergent challenges of Generative AI systems (HTI, 2024)

	Generative AI	Narrow or traditional AI
Purpose: Range of use	Generative AI systems are being deployed for a wide array of different tasks, but will not give accurate answers to queries outside their core training knowledge.	Narrow AI systems tend to be specifically trained and deployed within tight boundaries and controlled conditions.
Context: User familiarity and training	Generative AI systems are increasingly used by employees outside of technical teams, and require special training to be used appropriately. Employees may also be using these systems in ways unintended or permitted by their organisation (known as 'shadow' AI use).	Narrow AI systems tend to be designed and used by a small group of technical experts.
Data: Intellectual property	Systems often contain and may reproduce embedded, copyright-protected training data.	IP risks apply to training data, but input data to Narrow AI systems is often more knowable and manageable.
Data: Confidentiality and personally identifiable information	Systems may contain and reproduce confidential information or personally identifiable information including data inputted by users.	Data protection and confidentiality risks relate primarily to how data flows through the system.
Data: Quality and provenance	The scale of training data means quality can be dubious. Data quality can compromise fine-tuning. Low-quality prompts can compromise outputs.	Data provenance, quality and management are critical, but more knowable.
Data: Bias and fairness	Complex, real-world bias is often deeply embedded in models.	Bias is heavily reliant on data representativeness and algorithm choice.
Technical architecture: Potential for misuse	Deliberate, spontaneous, or user-induced production of harmful, misleading or manipulative content.	Misuse or misleading use of Narrow AI systems is possible, but easier to prevent.
Technical architecture: Accuracy	'Hallucinations' (namely, coherent responses that are false or incorrect) are common in Generative AI systems, particularly in response to prompts asking for information outside core training data.	The accuracy of Narrow AI systems can vary widely, but is generally more consistent and known.
Technical architecture: Reproducibility	Generative AI systems tend to produce very different results from similar input. Randomness is often deliberately used to improve output quality.	Narrow AI systems tend to be more predictable and stable, allowing for more systematic testing.

	Generative AI	Narrow or traditional AI
Technical architecture: Security	Generative AI systems are subject to 'prompt injection' (a type of cyber attack[8]) and other novel attacks, and may themselves be used to generate malicious code, as well as traditional cyber security risks.	Narrow AI systems are exposed to a wide range of traditional cyber security risks.
Technical architecture: Interpretability / Explainability	Generative AI systems feature complex models that are not interpretable. Their internal parameters do not support human understanding.	Most Narrow AI models allow for a level of interpretability, aiding accountability, and human learning from AI.
Automation: Automation bias	Users tend to trust AI systems more if they can communicate naturally. However, AI systems cannot express how confident they are in their responses.	Automation bias can also apply to Narrow AI systems, particularly if they are generally regarded as more reliable than humans. However, Narrow AI systems can be designed to provide a confidence level for a given output, which can offset this.

8 A 'prompt injection' is a type of cyberattack specifically aimed at General-purpose AI models that uses deceptive prompts to manipulate AI into leaking sensitive data or spreading misinformation. See IBM's What is a prompt injection attack? for more information



BOX 10: The challenge of explainability

AI systems are developed in very different ways to traditional software. Advanced AI systems are mainly developed through machine learning using historical data. General AI systems use variables known as 'weights' to make connections between various data points. 'Weights' determine the strength and nature of the connections between data points.

Simple AI models might have thousands of these weights, but larger and more complex models, like GPT-4, use billions of weights. By considering this larger set of variables the performance of these models is improved. However, it also decreases explainability. This makes it harder to explain how the AI model made its decisions, as well as to test the system and confirm its outputs.

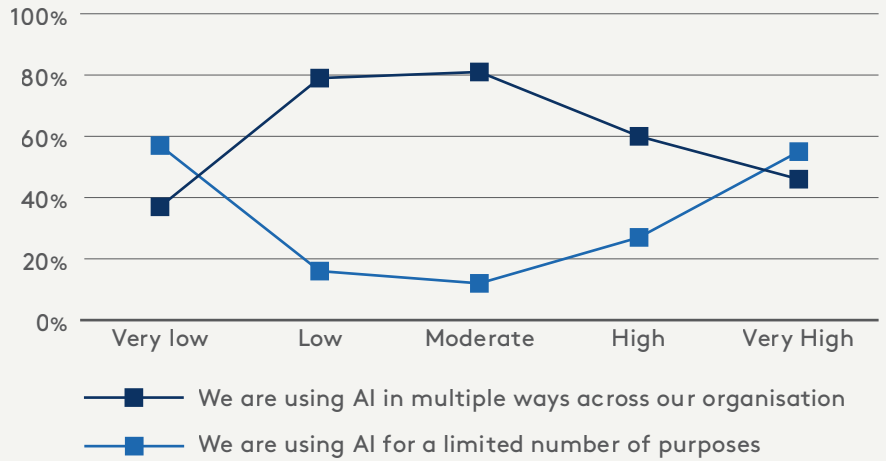
If a decision made by an AI system has legal or similarly significant effect, this lack of explainability can make it challenging for organisations to justify or sufficiently explain the reasons for the AI's output or decision.

2.4 PERCEPTIONS OF AI RISK AMONGST CORPORATE LEADERS

Perceptions of AI risk by business leaders, including directors, changes with experience (**Figure 3**).

Research conducted by HTI in 2022 found that corporate leaders who reported using AI in multiple ways across the organisation perceived AI risks as being either very low or very high. The inverse was true of leaders who reported limited use of AI within the organisation, with the majority of these leaders perceiving the risk as low-moderate.

FIGURE 3: Level of perceived risk posed by AI systems to organisations from directors and senior business leaders



These findings also suggest that organisations who are at the earlier stages of AI deployment may underestimate the potential risks and thus the need for governance transformation.

Conversely, corporate leaders with more AI experience perceive risk as clustered into two distinct categories: very low-risk systems (noting the tendency for leaders to be over-confident with regard to the risk of familiar AI use cases), and very high-risk systems (influenced by their exposure to observed negative outcomes from AI misuse and failure and the complexity of use).

Directors need to ensure their approach to AI balances the need to seize its opportunities with the need to mitigate AI's risks and harms. We discuss how directors can do so in A Director's Guide to AI Governance.

CHAPTER 3: Current obligations and the evolving regulatory landscape

KEY POINTS:

- Directors are ultimately responsible for the oversight of risk management throughout the organisation. This includes risk arising from AI.
- Existing legal obligations in the areas of privacy, consumer protection, intellectual property, cyber security, anti-discrimination, duty of care, and work, health and safety continue to apply, and will be relevant to AI use.
- The regulatory landscape is evolving rapidly. While regulatory approaches differ, a common set of safe and responsible AI principles underpin reforms locally and internationally.
- The Australian Government has committed to the introduction of a mix of mandatory and voluntary measures such as:
  - Consideration of the introduction of mandatory guardrails for AI deployment in high-risk settings.
  - Development of a voluntary risk-based AI Safety Standard.
  - Clarifying and strengthening existing laws to address AI risks and harms.

3.1 CURRENT LEGAL OBLIGATIONS IN RELATION TO THE USE OF AI

3.1.1 Directors' duties

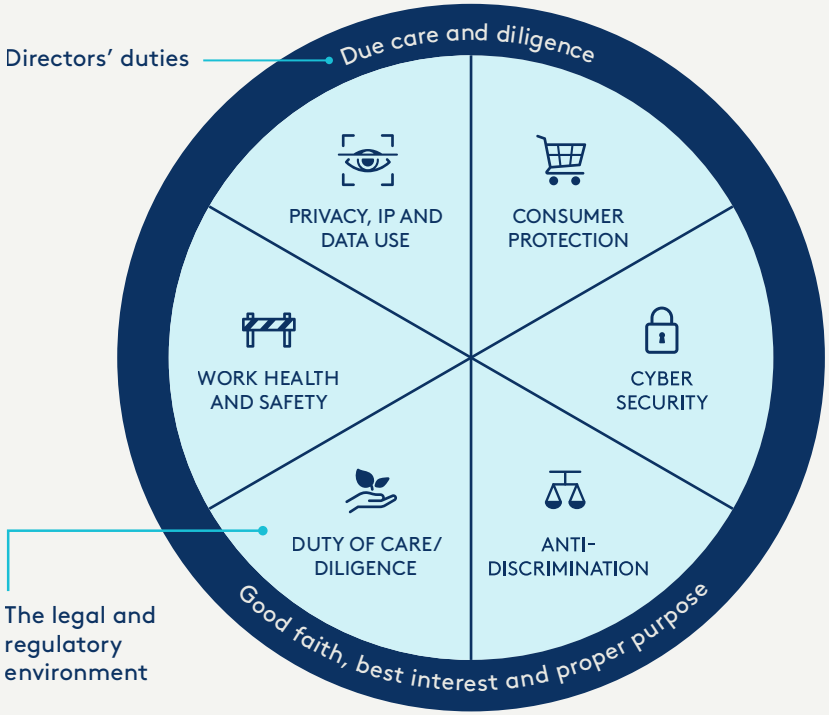
Directors are responsible for the oversight of the organisation's strategy and risk management processes. This includes managing AI risks and opportunities.

Directors have a fiduciary duty to act in the best interests of the company. When making decisions and providing oversight regarding their organisation's development and use of AI systems, directors are required to act:

- with due care and diligence; and
- in good faith, and for a proper purpose.

Focus is growing on the duty of care and diligence in the context of governance failures in meeting cyber security and privacy obligations, which are both relevant considerations in the use of AI systems. See **Box 12** for more information on key AICD cyber governance guidance.

FIGURE 3: Existing legal obligations when using AI



BOX 11: Directors' best interests' duty

In 2022, the AICD commissioned legal advice from Brett Walker AO SC and Gerald Ng of Counsel setting out their views on the content of the 'best interests' duty under section 181(1)(a) of the Corporations Act.

The opinion made clear that directors have considerable latitude in determining where the interests of the company lie, and that the law does not assume shareholder or member interests are best served by ignoring other stakeholders. The opinion confirms that corporate reputation is a legitimate director consideration, and there is no reason why directors could not have regard to the interests of customers, employees and the community more generally, provided that there is a rational justification for doing so by reference to the long-term interests of the company.

Drawing on this legal opinion, the AICD's Directors' 'best interests duty' Practice Statement states that, **as a guiding principle, directors should take a long-term view of where the company's interests lie.** Impact on customers and other stakeholders should be front of mind.

3.1.2 Existing legal obligations for organisations using AI

To discharge their duties, directors should understand the external legal and regulatory environment that applies to their company and its use of AI (see **Figure 3**).

While stand-alone AI regulation has not yet been introduced in Australia, a range of existing laws already apply to the design, development, and use of AI systems. The Australian Government has also foreshadowed further reform of these laws to apply more directly to AI use.

Some laws place obligations on the organisation, while others apply to directors and officers individually.

PRIVACY

Personally identifiable data is often collected and used to train and develop AI systems, or may be ingested in, or used by, an AI system. Organisations need to consider their privacy obligations pertaining to the collection, storage and use of personal information (see the ***Privacy Act 1988* (Cth)**). Significant changes to privacy law are being proposed under the **Privacy Act Review**.

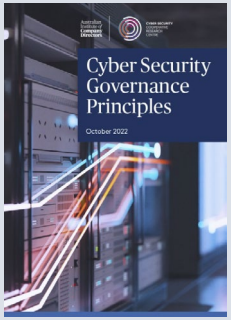
CYBER SECURITY

Cyber security is a key consideration for organisations developing and deploying AI, given AI's reliance on data and the increased risk of cyber security breaches. Depending on sector, risk management, notification and reporting obligations may apply to organisations and directors.

Directors should be aware of industry or sector-specific regulatory requirements that place obligations on how boards oversee digital, data and cyber security risks. These include risk management obligations under the **Security of Critical Infrastructure Act 2018 (Cth)** and requirements on the role of the financial services boards under Australian Prudential Regulation Authority prudential standards. Key requirements are summarised in the Cyber and Infrastructure Centre publication **Overview of Cyber Security Obligations for Corporate Leaders**.

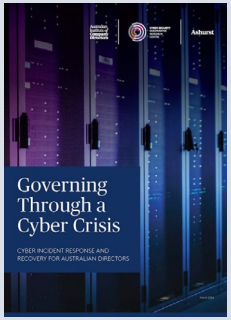
BOX 12: Director resources on cyber governance

The AICD and its partners have produced two major director resources on cyber security governance:



The AICD and Cyber Security Cooperative Research Centre (CSCRC)'s **Cyber Security Governance Principles** were launched in October 2022 to provide a governance framework for directors overseeing cyber security risk.

The Cyber Security Governance Principles have had over 25,000 downloads to date, and have been taken up by organisations locally and internationally.



In February 2024, the AICD and CSCRC partnered with Ashurst to publish a **'Governing through a Cyber Crisis'** to assist boards and directors in governing through a material cyber incident.

The resource expands on the Cyber Security Governance Principles and was informed by insight from senior Australian directors, cyber security advisors and Government.

CONSUMER PROTECTION

The provision of AI-enabled products or services which are used by organisations engaging with consumers – or are purchased and used by consumers directly – is subject to Australian Consumer Law (ACL). The ACL contains prohibitions against misleading or deceptive conduct, unconscionable conduct, and making false or misleading representations. Consumer warranties and guarantees (including that products are of an acceptable quality and are reasonably fit for purpose) and liability for harm caused by safety defects (e.g. where the organisation is a manufacturer under the ACL) are also relevant.

ANTI-DISCRIMINATION

The outputs of AI systems can directly or indirectly discriminate against individuals on the basis of protected attributes due to automated bias. Organisations have obligations under various anti-discrimination laws to prevent discrimination based on protected attributes such as a person's age, disability, disability carer status, race, colour, descent, national or ethnic origin, immigrant status, sexual orientation, gender identity, intersex status, marital or relationship status, pregnancy status, breastfeeding or family responsibilities.

DUTY OF CARE

Organisations may have a duty of care towards people who use or are impacted by an AI system. The law of negligence requires that where an organisation has a duty of care to a class of persons, the organisation must exercise the standard of reasonable care of a reasonable person in the circumstances to avoid foreseeable injury or loss. A failure to do so may mean that the organisation is liable for the loss or injury suffered by those persons to whom a duty is owed.

WORK HEALTH AND SAFETY

Deployment of AI systems within a workplace context can introduce risks of physical and psychological harm to employees. Directors must exercise due diligence in their oversight capacity so that organisations can meet their work health and safety (WHS) obligations.

INTELLECTUAL PROPERTY

AI systems, particularly Generative AI, are trained using large amounts of data, including written text, images, videos, or music. The unauthorised use of these works may have intellectual property implications for organisations, such as breaches of copyright. Conversely, copyright protection may not apply to works created by Generative AI systems.

BOX 13: AI and Human Rights

Privacy rights, non-discrimination and equality rights, freedom of expression, assembly and association and economic, social and cultural rights are among the myriad of rights that may be impacted by AI. In addition, the right to effective remedy may be undermined through the opaque design of 'black box' systems, failure to provide redress for those impacted by an AI system or product, or lack of awareness that an AI system was used.

Private corporate actors are expected to ensure that their services and products are not contributing to, or causing, human rights abuses, in line with the UN Guiding Principles on Business and Human Rights.

Conducting human rights impact assessments is necessary to:

- assist organisations to identify potential adverse effects from AI development;
- take action to remedy the risk of adverse effects;
- monitor mitigation measures; and
- communicate with impacted communities.

In addition, algorithmic impact assessments help organisations forecast negative impact and take steps to mitigate AI harms.

3.2 THE EVOLVING REGULATORY LANDSCAPE

3.2.1 International trends

Globally, jurisdictions are seeking to require those in the AI supply chain to identify, mitigate and monitor AI systems for risks and harms. However, approaches to the regulation of AI systems diverge.

PRESCRIPTIVE REGULATION

The European Union's *Artificial Intelligence Act* (EU AI Act) and Canada's proposed *Artificial Intelligence and Data Act* both adopt horizontal, economy-wide laws that regulate AI as a technology. These laws include requirements that will be overseen by new AI regulators.

China has also introduced laws which target particular techniques or applications such as Generative AI and 'deepfakes'.[9]

Importantly, approaches reflect the values of jurisdictions. For example, the core of the EU AI Act is the risk that AI systems and models pose to EU fundamental rights, whilst China's Generative AI regulations require providers of Generative AI systems to ensure generated content upholds core socialist values.

REGULATORY GUIDANCE AND VOLUNTARY STANDARDS

The UK and Singapore rely more heavily on regulatory guidance. The UK recently tasked existing sectoral regulators to issue guidance with reference to a common set of AI principles, while Singapore has released a series of voluntary tools to assist in the implementation of AI.

Voluntary technical standards and frameworks are also being produced through bodies such as the International Organisation for Standardisation (ISO) and US National institute of Standards Technology (NIST) to assist organisations use AI in line with emerging best practice.

AI developers, such as Microsoft and Google, have also made public their own AI governance and risk management frameworks while also making voluntary commitments to promote the responsible development of AI (see Microsoft's AI governance approach in **Case Study 4** in A Director's Guide to AI Governance).

Directors should be particularly aware of ISO/IEC 42001:2023, an international standard that sets out a structured approach for organisations to address the unique challenges of AI systems and manage the related risks and opportunities.

9 A deepfake is a digital photo, video or sound file of a real person that has been edited or manipulated to create a realistic but false depiction of them.

COMMON THREADS IN THE REGULATORY APPROACH TO AI

While jurisdictions may differ in regulatory approach and enforcement, important commonalities include:

- Being underpinned by a set of **safe and responsible AI principles** (see **section 3.3**).
- Defining AI with close reference to the OECD definition, being: *"a machine based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."*[10]
- Risk-based approaches (see **Box 14** and **Box 15**).
- Incentivising organisational governance through technical standards and other measures, such as requirements at the design and development stages. This includes testing requirements and risk assessments to prevent AI-facilitated harms from arising.
- Added focus on highly capable general-purpose models (General AI).

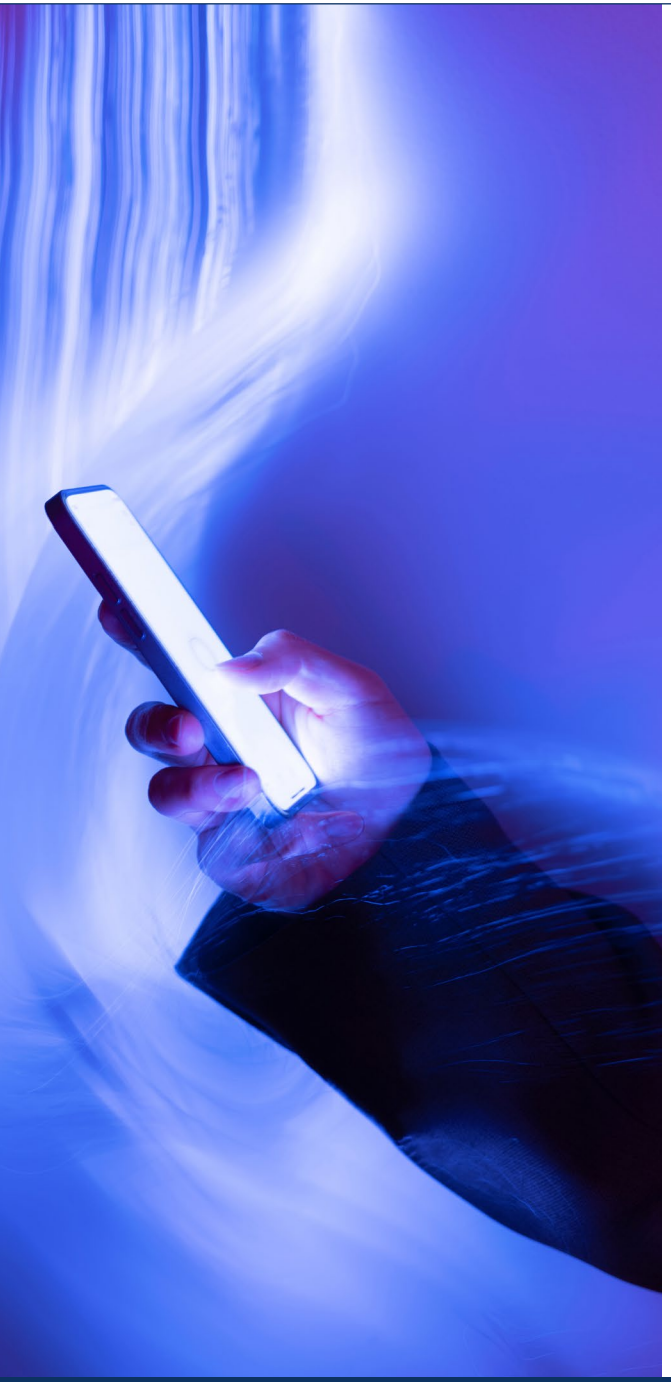
BOX 14: What is high-risk AI?

There is not yet a commonly agreed definition of high-risk AI uses or settings. This varies substantially by jurisdiction. Examples include:

- **EU:** certain products or safety components of products such as medical devices, machinery, toys, lifts, aircraft; as well as AI systems used in biometrics, critical infrastructure, education, access to essential services (public and private), law enforcement, immigration, administration of justice and democratic processes.
- **Canada:** screening systems impacting access to services or employment; biometric systems used for identification and inference; systems that can influence human behaviour at scale; and systems critical to health and safety.

In addition to these common themes, a number of commitments at bilateral, regional and multilateral levels have been made to collaborate and cooperate on AI. The Bletchley Declaration, which has been signed by the EU and 28 other countries including Australia, committed signatories to the sharing of knowledge on AI risk and safety and governance approaches, assurance techniques and global technical standards. Commitments were also made by G7 nations pursuant to the Hiroshima AI Process, and the Global Partnership on AI.

10 Stuart Russell, Karine Perset and Marko Grobelnik, Updates to the OECD's definition of an AI system explained, OECD.AI Policy Observatory (Blog post, 29 November 2023).



BOX 15: The EU's risk-based approach to AI regulation

In December 2023, the EU reached a landmark provisional agreement on the world's first comprehensive law on AI, following protracted negotiations. The EU AI Act takes a risk-based approach to the regulation of AI by applying four categories of risks:

- **Minimal or no risks:** Uses in this category can continue unimpeded (although voluntary codes are encouraged). Examples of AI falling in this category include AI-enabled recommender systems or spam filters.
- **Limited risks:** Uses in this category can continue, but are subject to some light transparency obligations.[11] Examples of AI falling in this category include chatbots.
- **High risks:** This includes AI used in the following eight contexts: biometrics, critical infrastructure, education, employment, access to public and private essential services, law enforcement, immigration and the administration of justice. AI use in these contexts will be allowed, but subject to stringent requirements.[12] A number of exceptions apply, including where the AI system falls within one of these high-risk categories but does not pose a *"significant risk of harm to the health, safety or fundamental right of natural persons including by not materially influencing the outcome of decision making"* because it satisfies one of four specific criteria.[13]
- **Unacceptable risks:** These uses, which are banned with limited exceptions, include cognitive manipulation, certain applications of predictive policing, emotional recognition in workplace and schools, social scoring and certain remote biometric identification system risk.

The EU AI Act was approved on 21 May 2024. It will enter into force 20 days after publication in the Official Journal of the EU, which is expected to be in June 2024. Most of the significant obligations under the Act, such as the obligations for high-risk AI use, will come into force 2 years after the Act's commencement.

11 Including informing users that they are interacting with an AI system and marking synthetic audio, video, text and images content as artificially generated or manipulated for users and in a machine-readable format.

12 Including comprehensive mandatory compliance obligations in respect of risk mitigation, data governance, detailed documentation, human oversight, transparency, robustness, accuracy and cybersecurity. Entities using high-risk AI will also be required to undertake "conformity assessments" to evaluate and confirm compliance with the EU AI Act as well as Fundamental Rights Impact Assessments. Impacted persons will also have a right to launch complaints about AI systems and to receive explanations about decisions based on high-risk AI systems that impact their human rights.

13 Being: (1) the AI system is intended to perform a narrow procedural task; (2) the AI system is intended to improve the result of a previously completed human activity; (3) the AI system is intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment without proper human review; or (4) the AI system is intended to perform a preparatory task to an assessment relevant for the purpose of the cases (otherwise listed as potential high-risk AI uses).

3.2.2 Australian regulatory and policy developments

Australian regulators are increasingly focused on the application of existing laws to the use of AI and algorithms by organisations. Notable regulatory investigations and/or enforcement actions include the Office of the Australian Information Commissioner's finalised investigations into **Clearview AI**, the Australian Competition and Consumer Commission (ACCC)'s proceedings against **Trivago** (see **Box 16**), and the Australian Securities and Investment Commission (ASIC)'s proceedings against Insurance Australia Limited (see judgment **here**).

Whilst there has been increased focus on enforcement, there is also some recognition of the limitations of existing regulatory guardrails in Australia to address the risks posed by AI.

BOX 16: Case spotlight: *ACCC v Trivago*

In January 2020, in proceedings initiated by the ACCC, the Federal Court of Australia found that Trivago N.V. had breached the Australian Consumer Law by misleading consumers when claiming that its website provided customers with the best deal available for a given hotel.

This finding was made in circumstances where, in determining which rooms to highlight, the algorithm powering Trivago's website placed significant weight on which online hotel booking site paid Trivago the highest fee. As a result, two-thirds of the time it did not highlight the cheapest rates for consumers. Trivago admitted that this amounted to approximately $58 million in fees from offers that were not the cheapest available to customers, causing consumers to overpay for hotel rooms by around $38 million.

The Federal Court ordered Trivago to pay penalties of $44.7 million for its misleading representations about hotel room rates made on its website and in television ads, a decision upheld by the Full Federal Court on appeal.

This case reinforces that AI-powered algorithms are subject to the same consumer laws as any other business process.

In January 2024, the Australian Government released its interim response to its discussion paper on supporting responsible AI. The response sets out the federal government's agenda for AI regulation in the context of broader law reform relevant to AI, such as privacy reform and online safety reform.

Reforms flagged by the Government include:

- Consideration of the introduction of mandatory guardrails for AI deployment in high-risk settings;
- Development of a voluntary risk-based AI Safety Standard;
- Consideration of labelling and watermarking of AI in high-risk settings;
- Clarifying and strengthening existing laws to address AI harms and risks; and
- Supporting international engagement on AI governance and ensuring interoperability with Australian responses.

Australian organisations and directors should expect increased scrutiny by regulators and additional risk management and governance requirements associated with AI system use.

> "Our job is to mitigate the known risks - and, in doing so, bend the trajectory away from the worst imagined outcomes, so that they never materialise."

Source: Joe Longo, ASIC Chair
ASIC x UTS: AI Regulators' Symposium – 21 May 2024

3.3 SAFE AND RESPONSIBLE AI PRINCIPLES

Both international and Australian approaches to the regulation of AI are founded on the need to ensure that AI use is safe and responsible.

Safe and responsible AI is generally considered by reference to overarching principles including security, safety, fairness, accountability, transparency and explainability, and redress.

While many permutations of such principles exist, the two most relevant for Australian organisations are the OECD AI Principles, and Australia's AI Ethics Principles.

Australia's AI Ethics Principles largely aligns with the OECD principles and provides a voluntary framework for businesses and governments. The key principles underpinning the framework include:

- **Human, societal and environmental well-being:** AI systems should benefit individuals, society and the environment.
- **Human-centred values:** AI systems should respect human rights, diversity, and the autonomy of individuals.
- **Fairness:** AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.
- **Privacy protection and security:** AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.
- **Reliability and safety:** AI systems should reliably operate in accordance with their intended purpose.
- **Transparency and explainability:** There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.
- **Contestability:** When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.
- **Accountability:** People responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

For guidance on boardroom ethical decision-making more broadly, see the AICD and Ethics Centre's Ethics in the Boardroom resource.

Where to from here? Governance implications

For practical guidance on how to implement safe and responsible AI governance, see Part 2 of this series, '<u>A Director's Guide to AI Governance</u>'.

Acknowledgements

The AICD would like to acknowledge our Guide co-authors:

Professor Nicholas Davis MAICD
Co-Director, Human Technology Institute
University of Technology Sydney

Lauren Solomon
Lead, AI Governance (until April 2024)
Human Technology Institute,
University of Technology Sydney

Llewellyn Spink
AI Corporate Governance Specialist
Human Technology Institute,
University of Technology Sydney

The AICD would also like to acknowledge and thank the following people who were involved in the review of the Resource:

- Alison Kitchen MAICD
- Kee Wong FAICD
- Phil Coffey GAICD
- Wendy Stops GAICD

ABOUT AICD

The Australian Institute of Company Directors is committed to strengthening society through world-class governance. We aim to be the independent and trusted voice of governance, building the capability of a community of leaders for the benefit of society. Our membership includes directors and senior leaders from business, government and the not-for-profit sectors.

ABOUT HTI

The UTS Human Technology Institute (HTI) is an impact-oriented institute building human values into new technologies. Bringing together policy, legal and technical experts, HTI provides independent expert advice, policy development, capability building, and data science solutions to support government, industry and civil society.

DISCLAIMER

The utmost care has been taken to ensure this document accurately reflects the legislative and regulatory landscape as at the date of publication. However, this is an area subject to constant regulatory and legal change. The material in this publication does not constitute legal, accounting or other professional advice. While reasonable care has been taken in its preparation,the AICD and HTI do not make any express or implied representations or warranties as to the completeness, reliability or accuracy of the material in this publication. This publication should not be used or relied upon as a substitute for professional advice or as a basis for formulating business decisions. To the extent permitted by law, the AICD and HTI exclude all liability for any loss or damage arising out of the use of the material in the publication. Any links to third party websites are provided for convenience only and do not represent endorsement, sponsorship or approval of those third parties, any products and services offered by third parties, or as to the accuracy or currency of the information included in third party websites. The opinions of those quoted do not necessarily represent the view of the AICD or HTI. The AICD and HTI reserve the right to make changes without notice where necessary.

Copyright

For more information about A Director's Introduction to AI:

T: 1300 739 119

E: **policy@aicd.com.au**

JOIN OUR SOCIAL COMMUNITY

aicd.com.au

CCT1403_24