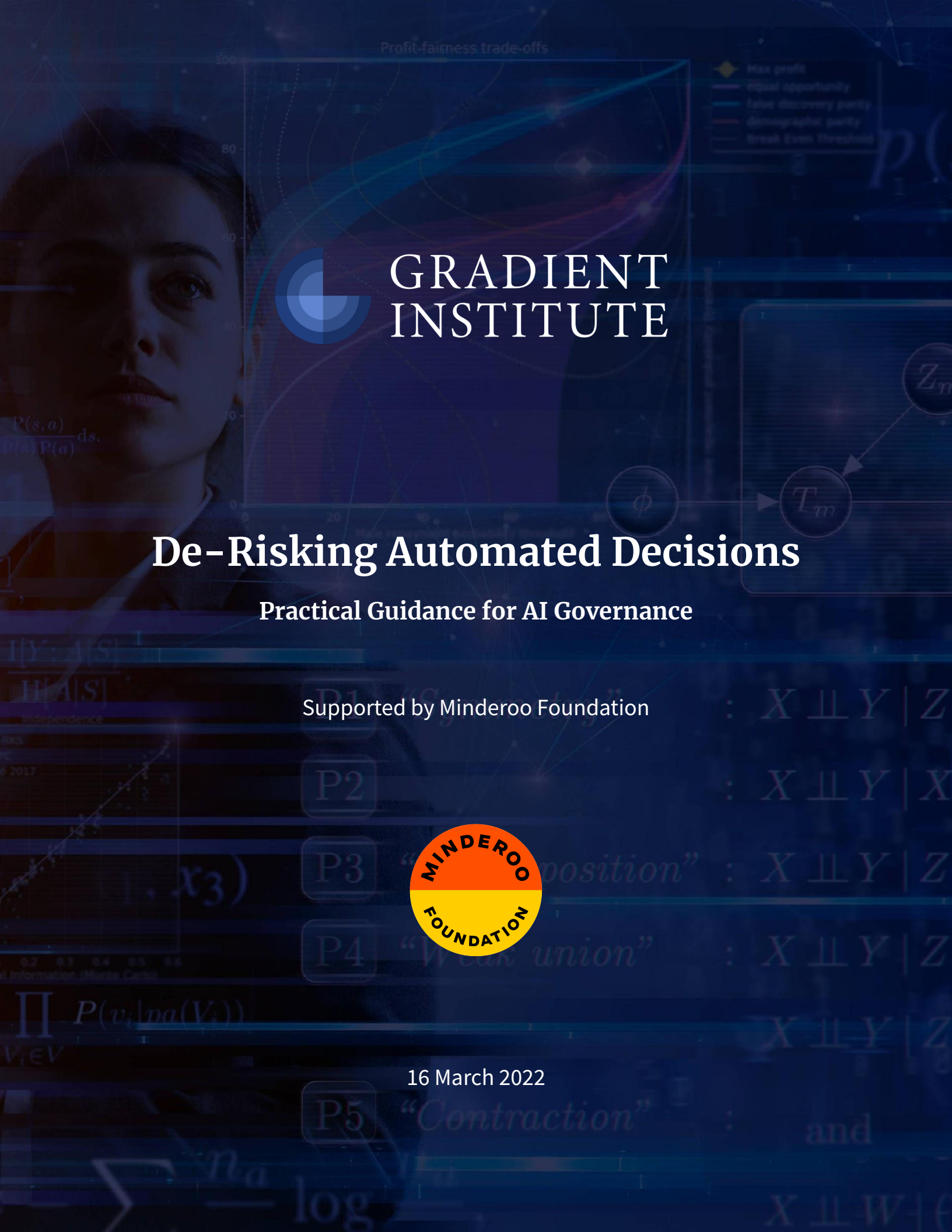


GRADIENT INSTITUTE

De-Risking Automated Decisions

Practical Guidance for AI Governance

Supported by Minderoo Foundation

MINDEROO FOUNDATION

16 March 2022

This report has been prepared by Gradient Institute, and partially supported by Minderoo Foundation.

Authors

Tiberio Caetano, Jenny Davis, Chris Dolman, Simon O'Callaghan, Kimberlee Weatherall

Acknowledgements

The authors would like to acknowledge

for detailed feedback, discussions and contributions to specific sections of this report: José-Miguel Bello y Villarino, Lachlan McCalman, Alistair Reid

for discussions and detailed feedback on earlier versions of this report: Claire Benn, Jake Blight, Nicky Burns, Kobi Leins, Emma McDonald, Stuart Powell, Linda Przhedetsky, Kashif Qadir, Dan Steinberg, Barry Wang

for discussions and input throughout the development of this report: Zena Assaad, Will Bateman, Henry Fraser, Nicholas Gruen, Ian Opperman, Steve Ramm, Melanie Trezise, Maciej Trzaskowski, Phillip Ward, Elizabeth Williams, Lexing Xie

for regular discussions that helped shape the ideas reflected in this report, and for detailed feedback on earlier versions of this report: Rachel Howard, Bill Simpson-Young.

Contact

Gradient Institute Ltd.
https://gradientinstitute.org
info@gradientinstitute.org

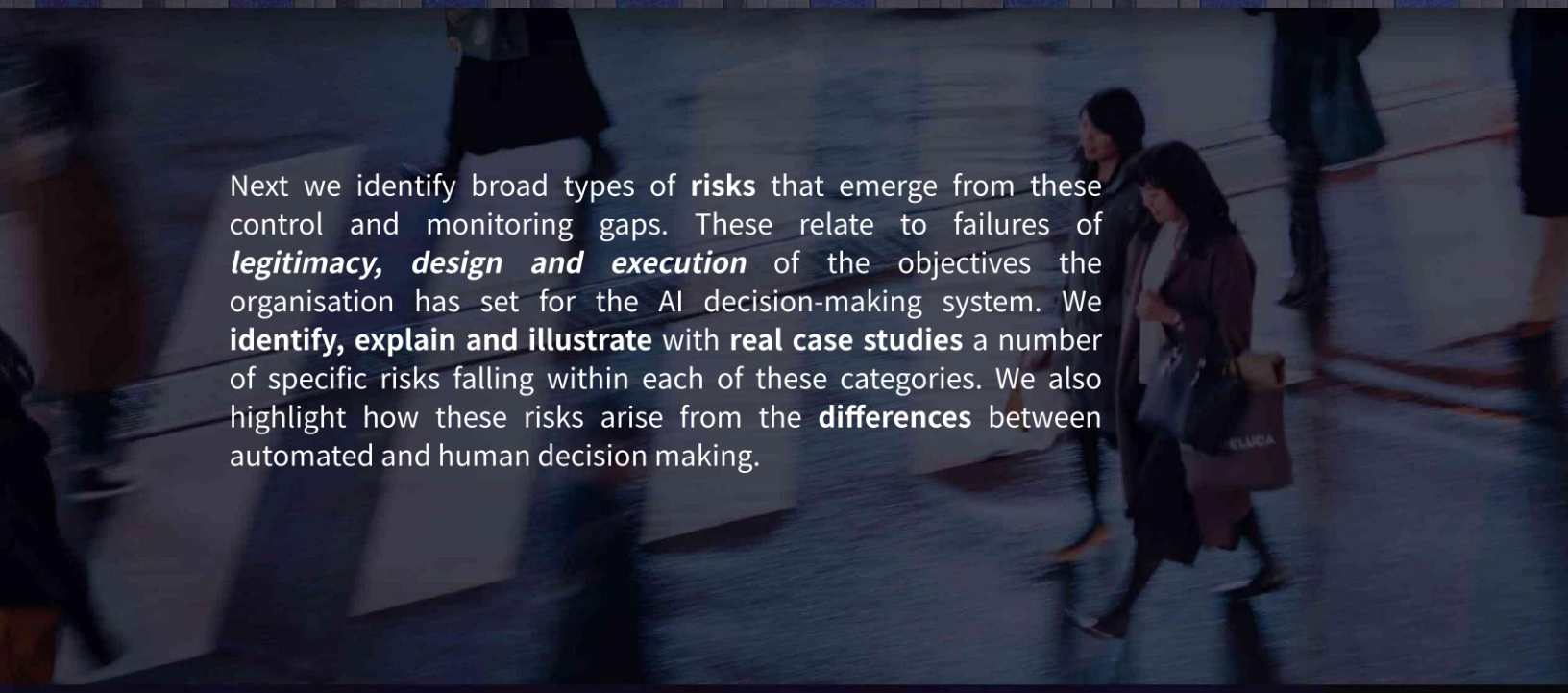

Z
T
Y
Executive Summary

The growing use of AI is challenging traditional methods of governance. AI is increasingly being used for automated decision making: from credit scoring and insurance pricing to marketing, health triage, recruitment and provision of government services. But when decisions are made by 'AI Systems' rather than humans, traditional governance methods of control and monitoring – designed for human decision makers, not machines – can, and do, fail.

Current governance structures have failed to control AI. As illustrated in this report, *novel risks* created by the use of AI systems for decision making have resulted in unlawful, discriminatory, opaque and unaccountable decisions. The humans responsible for the design, deployment and use of these systems are invariably surprised by these failings.

This report provides **general guidance on de-risking automated decisions** in light of these growing risks. How does traditional governance break down when important decisions are made by AI systems, and what can be done about it? That's the question we address.

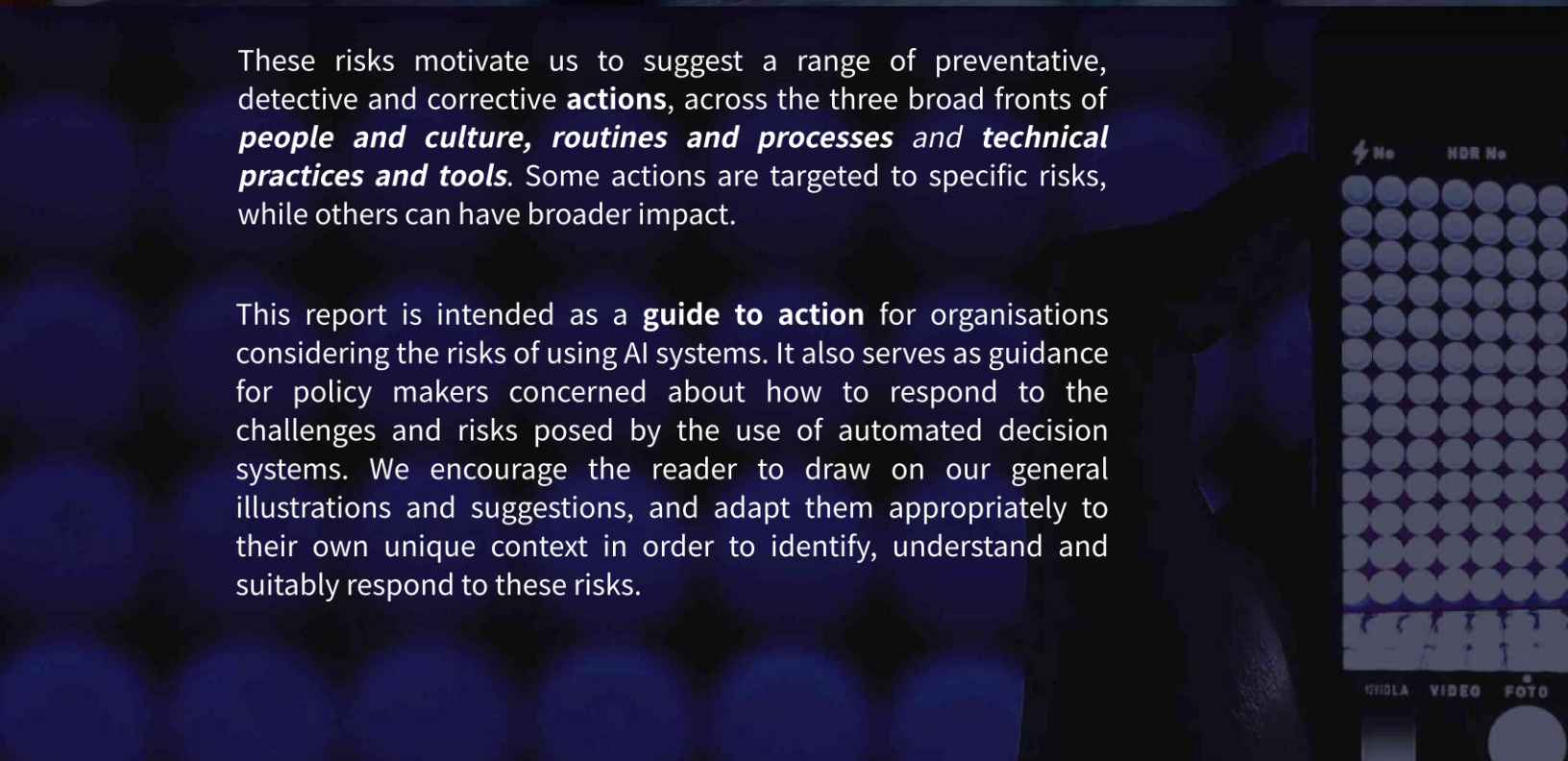
First we explain how the differences in decision making between people and machines create **fundamental control and monitoring gaps** when assigning decisions to an AI system. These gaps emerge due to a number of factors, such as the challenges of equipping an AI system with elements of common sense, or a basic understanding of moral concepts, or an ability to explain its inner workings back to humans.

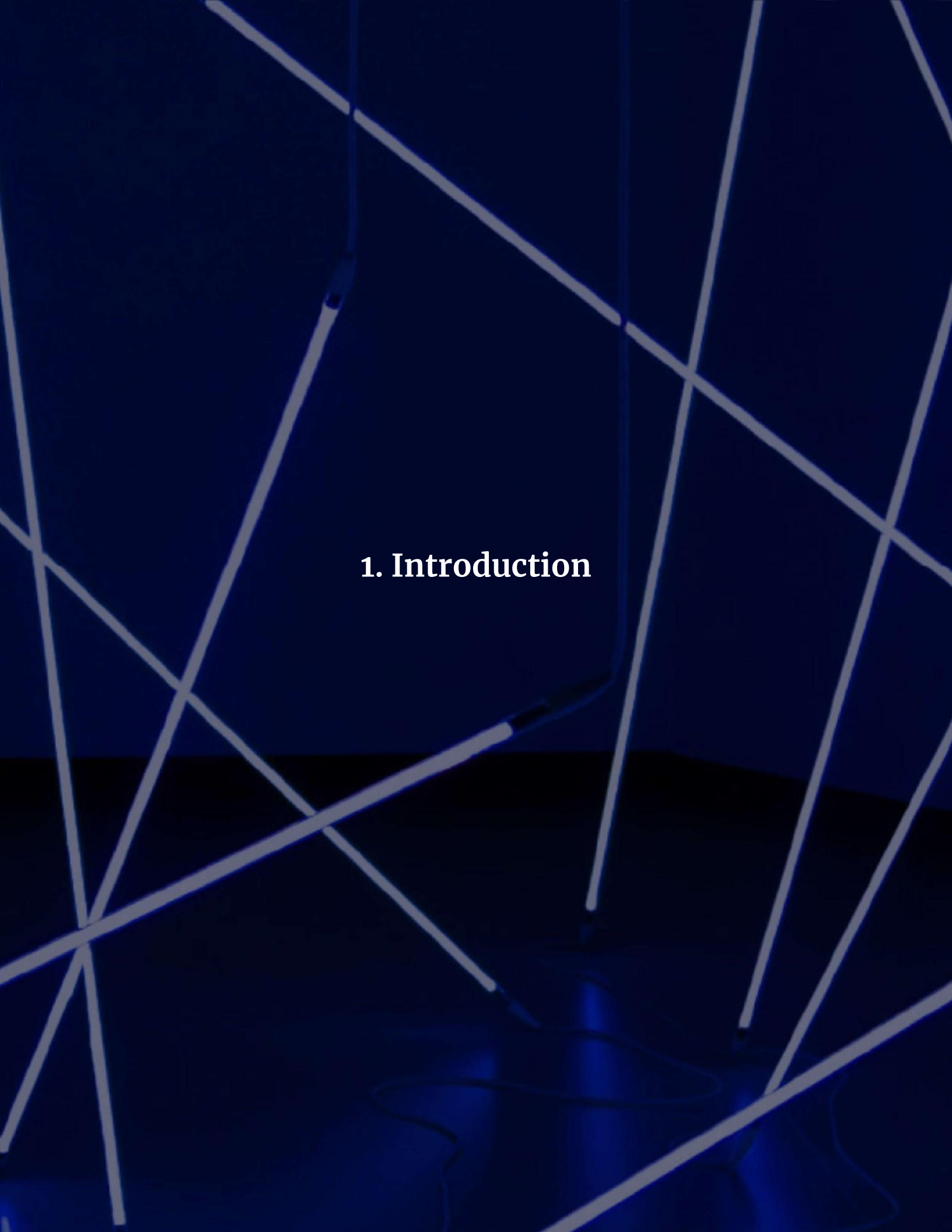


Next we identify broad types of **risks** that emerge from these control and monitoring gaps. These relate to failures of ***legitimacy, design and execution*** of the objectives the organisation has set for the AI decision-making system. We **identify, explain and illustrate** with **real case studies** a number of specific risks falling within each of these categories. We also highlight how these risks arise from the **differences** between automated and human decision making.

These risks motivate us to suggest a range of preventative, detective and corrective **actions**, across the three broad fronts of ***people and culture, routines and processes*** *and* ***technical practices and tools***. Some actions are targeted to specific risks, while others can have broader impact.

This report is intended as a **guide to action** for organisations considering the risks of using AI systems. It also serves as guidance for policy makers concerned about how to respond to the challenges and risks posed by the use of automated decision systems. We encourage the reader to draw on our general illustrations and suggestions, and adapt them appropriately to their own unique context in order to identify, understand and suitably respond to these risks.





1. Introduction

1.1 Context and Background

Organisations are increasingly delegating their decisions to algorithms, or "AI systems".[1] These systems are used today to decide who gets insurance, who gets a loan and who gets a job. They curate people's information diet via personalised news feeds, create parole and sentencing risk scores, administer social services systems, choose the ads people see, suggest traffic routes and price insurance policies. The list keeps growing by the day as organisations realise that using AI can lead to faster, more accurate, and more "profitable" decisions.

Unfortunately, as we show in this report, there is now overwhelming evidence that the use of AI for decision-making has the potential to produce unlawful, immoral, discriminatory outcomes for individuals through opaque and unaccountable decision processes. This has ignited a range of worldwide policy responses, including concrete regulatory proposals, such as the European Union's proposal for an 'Artificial Intelligence Act'.[2] The EU proposal is the first significant legal framework specifically designed for regulating the use of AI: it adopts a risk-based approach to AI regulation, prohibiting certain uses of AI and proposing extensive provisions for others identified as "high-risk". Likewise, Brazil has a new proposal for AI regulation which is also risk-based.[3] In Australia, the Australian Human Rights Commission report on Technology and Human Rights made a number of policy recommendations for AI regulation,[4] also with a focus on the responsible use of AI through a risk-aware perspective, for example through proactive impact assessments.[5]

These harms are arising from unwarranted trust in (or at least reliance on) AI. Humans and machines make decisions differently. While humans have common sense and are able to navigate different contexts with ease, machines have no built-in moral judgement and only perform well in relatively narrow domains. On the other hand, where humans can be susceptible to opinions, emotions, cognitive biases, fatigue and self-interest, machines will do

---

[1] We do not mean delegation in a legal sense, in which a delegate is given the *legal* authority of the original decision maker (since algorithms are not legal subjects). Rather we refer to the fact that in modern Artificial Intelligence (AI) systems, humans specify the high-level objective to be pursued by the system, as well as the data to be used, and the AI system automatically generates from those high-level parameters outputs such as predictions, decisions, recommendations or actions.

[2] *Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (Data Governance Act), Brussels,* 25.11.2020 COM(2020) 767 Final, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

[3] https://www.camara.leg.br/noticias/811702-camara-aprova-projeto-que-regulamenta-uso-da-inteligencia-artificial

[4] Australian Human Rights Commission, *Human Rights and Technology Final Report* (Australian Human Rights Commission, 2021) available at https://tech.humanrights.gov.au/downloads

[5] E. Moss, E. Watkins, R. Singh, M. C. Elish and J. Metcalf, Assembling Accountability: Algorithmic Impact Assessment for the Public Interest (June 29, 2021). Available at SSRN: https://ssrn.com/abstract=3877437

exactly as instructed, are reconfigurable, work relentlessly, can operate at a scale no human ever could, and will never complain.

These differences suggest human and automated decision-making systems fail in different ways. Governance of automated decisions needs to be designed to address these different categories of failure. **If decisions that were previously made by humans start to be made by machines, this poses a risk if governance systems fail to adapt**. **How does traditional governance break down when important decisions are delegated to machines, and what can be done about it?**

This report seeks to address this question. It explains and illustrates how AI decision-making challenges traditional governance and provides guidance on steps organisations can take to start addressing the challenges.

1.2 Our Contribution

Our purpose is to provide **general, practical guidance** for organisations on reducing risks[6] introduced by automated decision systems. Our contribution can be summarised as follows:

- We **identify** a non-exhaustive list of **types of risks** introduced or amplified by AI-driven decision making. We describe them in general terms, allowing the reader to adapt them to suit the specifics of their organisational or individual context. Through case studies, we illustrate how these risks, if unmitigated, result in avoidable harms.
- We **suggest** a non-exhaustive list of **interventions** to address the identified risks. Again we describe these interventions in general terms, so that they can be adapted suitably to the nature, scale and needs of an individual organisation. They include a range of preventative, detective and corrective measures.

Importantly, we do not attempt to provide any comprehensive "framework" for AI risk management or governance. Such efforts are being conducted at a global scale by standards bodies such as ISO[7] and NIST.[8] Our intent is to provide practical suggestions inspired by

[6]Acknowledging the standard definition of risk as the effect of uncertainty on objectives, in this report we are concerned with downside risks. Hence when using the term risk, this should be taken to mean an exposure to a potential harm or loss, rather than the potential for gain. This matches the colloquial discussion of AI risks, which focuses on downsides. Comparative terms such as 'high-risk', 'low risk' or 'increased risk' should be read as considering the expected loss arising from those risk exposures, rather than other potential measures of risk which could be considered.

[7] ISO is currently developing both a risk management standard (https://www.iso.org/standard/77304.html) and a management systems standard (https://www.iso.org/standard/81230.html) dedicated to AI.

[8] NIST (the US National Institute of Standards and Technology) has proposed an approach for identifying and managing biases in AI, based on risk: Schwartz, Reva et al, *A Proposal for Identifying and Managing Bias in Artificial Intelligence* (National Institute of Standards and Technology, 22 June 2021) https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270-draft.pdf

real-world failings which can be applied by organisations today. This is intended to complement, not replace, risk management frameworks and standards which are largely still under development, and which may yet take some time to be broadly accepted and adopted.

In this report, we consider risks of AI systems in general. At this level of generality, we do not believe it is appropriate to measure, rank or compare the magnitude of risks identified. Many risks will be irrelevant in some contexts, but critically important in others, meaning an overall ranking system may be misleading. We believe an organisation and its leadership possess the best specific knowledge to be able to judge which risks are most important for any particular context they may be considering. We do, however, make the following observations which should be appraised by organisations wanting to measure or prioritise AI risks or interventions in a specific context:

- **AI is an immature technology. Before asking how it can be used for a particular application, one should ask whether it should be used in that context at all**. Not only is AI *not* a panacea, but it is often the very cause of problems. This is usually due to a combination of its immaturity as a technology and the misplaced trust people tend to confer in it. In Appendices A and B we present questionnaires to help organisations assess whether the use of AI in a certain application is likely to be a bad idea. These are only general guiding tools and cannot replace a case-by-case, rigorous legal, ethical and accountable analysis by the organisation.
- **Many of the risks we identify are unintended**. They are typically a byproduct of ineffective governance of AI decision making, rather than risks consciously taken in pursuit of a potential reward. Organisations should have little tolerance for such risks and should seek to reduce them as much as possible where practical.
- **The process of quantifying AI risk is not mature, hence it is likely to be applied inconsistently and will be prone to error**. Risk quantification from the introduction of a new technology is likely to be extremely uncertain.[9] This uncertainty is especially pronounced for general-purpose technologies, which is the case of AI.[10]
- **When there is uncertainty or room for discretion in risk assessment, organisations have incentives to underestimate risk**. Risks are likely to be underestimated (or at least under-reported) if there are tacit incentives to under-report risk – a reality in many organisations. This became clear in our discussions with risk and compliance industry practitioners. While this issue is not specific to AI, the general purpose nature of this technology adds uncertainty and makes this issue more important. Risk

[9] R. C. Williamson, M. N. Raghnaill, K. Douglas and D. Sanchez, Technology and Australia's future: New technologies and their role in Australia's security, cultural, democratic, social and economic systems, Australian Council of Learned Academies, September 2015, www.acola.org.au

[10] Nicholas Crafts, Artificial intelligence as a general-purpose technology: an historical perspective, Oxford Review of Economic Policy, Volume 37, Issue 3, Autumn 2021, Pages 521–536, https://doi.org/10.1093/oxrep/grab012

managers should very deliberately acknowledge such matters when considering AI risks.

In summary, there are significant risks associated with AI systems. Standards and frameworks are in development at a global level to help organisations manage these risks, but organisations do not need to wait – they can act now to start reducing these risks. This report is intended to help organisations do just that.

1.3 Scope

An important difference between human and AI decision making is the role of data. Human decisions need not use any digital data at all, while AI systems require it. Many examinations of AI risks, therefore, focus on well trodden themes of data governance: privacy, security, consent, data quality and related topics. This report makes no attempt to add to this rich discussion, **focussing instead on the manner of automated decision making**, and the way it challenges core concepts of organisational governance.

We focus specifically on risks introduced or amplified by automated decision systems such as those currently in use in both the private and public sectors. Systems in scope are those producing decisions, predictions, recommendations or actions in which automation plays key roles that would otherwise have been performed by humans. Examples include marketing automation, credit scoring (or any other risk scoring), propensity scoring, personalised pricing, job applicant scoring, and more generally any type of automated recommendation or scoring. It can also include systems in which intermediate instead of final decisions are automated, such as loan approvals and risk audits.

Although we consider automated decision making in its general form, the report has an emphasis on its use within a corporate context. For an emphasis on administrative decision making, we refer the reader to earlier reports by the Commonwealth Ombudsman[11], the New South Wales Ombudsman[12] and the Attorney-General's Department.[13]

[11] Commonwealth Ombudsman, *Automated Decision-Making Better Practice Guide,* https://www.ombudsman.gov.au/publications/better-practice-guides/automated-decision-guide
[12] The new machinery of government: using machine technology in administrative decision-making. https://www.ombo.nsw.gov.au/news-and-publications/publications/reports/state-and-local-government/the-new-machinery-of-government-using-machine-technology-in-administrative-decision-making
[13] Administrative Review Council, *Report 46 - Automated Assistance in Administrative Decision-Making* 2004, https://www.ag.gov.au/legal-system/publications/report-46-automated-assistance-administrative-decision-making-2004

2. Concepts

2.1 AI Decisions

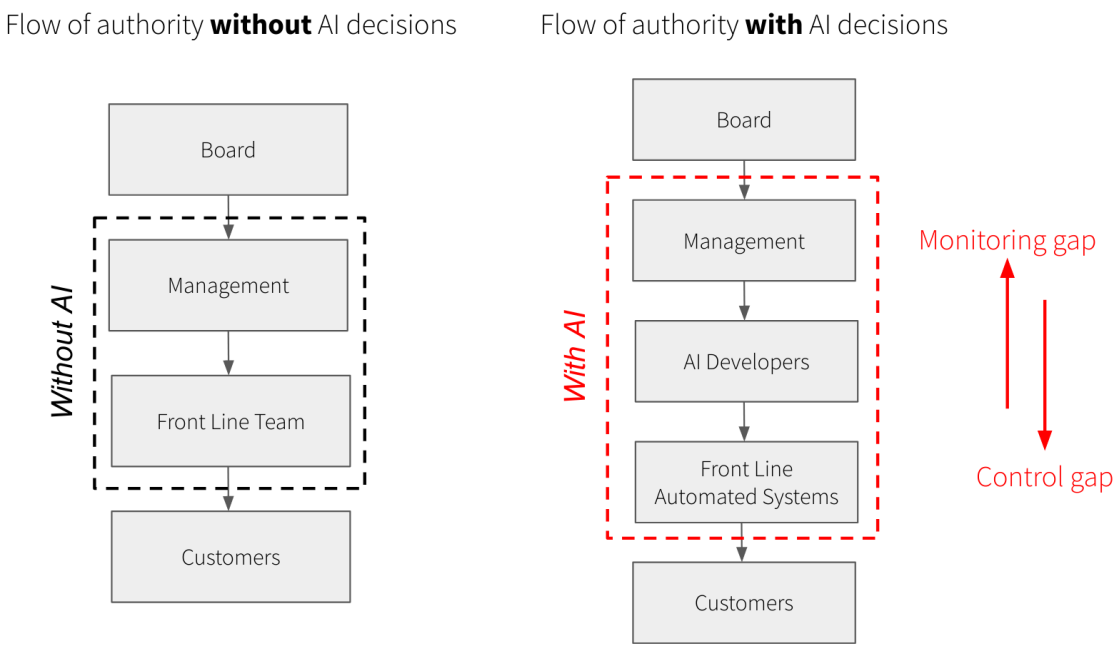
Consider an AI-driven loan application approval system. A bank has credit card transaction **data** and credit history data on its customers. The bank wants to know which new loan applications should be approved to deliver on the business **objectives** (e.g. maximising profit in the long term). An AI-driven solution to this problem might proceed along the following lines. AI developers use a "machine learning algorithm" to analyse the transactional data of customers that did and did not repay their loans. The algorithm learns which patterns of transactions correspond to a person being more or less likely to repay. The output of the machine learning algorithm is *another* algorithm: an "AI model". The AI model takes historical transaction data for a customer as input and produces as output the probability the customer will repay a loan in the future. The bank then applies this model to new applicants to estimate their repayment probability. From that probability (and any other relevant considerations), the system determines when to give out loans in order to achieve the bank's business objectives. The complete system is called an "AI system" for decision-making.[14]

The example outlined above can easily be generalised and adapted to many other settings. **The data and objectives can change, but the template is the same**. Since organisations have increasing accessibility to data and machine learning tools, the methodology is not only **conceptually general** but also **readily implementable**. This is what's facilitating AI adoption across a diverse range of sectors and settings – there are few (and diminishing) areas of society without the potential for AI adoption along the lines of the example above.

2.2 Control and Monitoring Gaps

What changes when we introduce AI decision-making into an organisation? A starting point is to examine where AI decisions fit within an organisation's formal flow of authority.

[14] Definitions here are difficult, and language is imprecise. In this report, focused on high level guidance, we can afford some level of ambiguity: we use "automated decision systems" and "AI Systems" interchangeably, as well as "AI decisions" and "automated decisions", or "automated decision making" and "AI decision making". As will be clear from the text, nevertheless, our focus is on forms of automated decision making that create different challenges for traditional governance mechanisms within organisations.



2.2.1 Flow of Authority without AI Decisions

In this scenario (depicted above left), all decisions impacting customers follow a human-to-human chain of authority. The board sets the strategy and high level policies for the organisation and delegates key functions to management. Management establishes business objectives based on the strategy; it also sets detailed rules, policies and procedures and delegates their execution to its front-line team, as well as the authority to exercise discretion in situations in which the rules are silent or unclear. The front-line team interacts with customers by following the rules and using discretion when required.

Failures can happen at different stages of this decision-making pipeline. The strategy itself may have issues; or the rules created by management to implement the strategy may be problematic; or the front-line team may not appropriately follow the rules (when those are right); or the use of discretion may be inappropriate.

Traditional management of these failure modes is done within the "three lines of defence model".[15] The business function (line one) provides customer service and manages risks. A separate risk and compliance function (line two) provides support, typically including advising on rules, checking they are followed and that the use of discretion is reasonable. The internal audit function (line three) reports directly to the board and provides an independent assessment on the activities and achievements of management, including whether the

[15] Risk and compliance: rethinking the three lines of defence. https://aicd.companydirectors.com.au/membership/company-director-magazine/2020-back-editions/november/risk-and-compliance-rethinking-the-three-lines-of-defence

business objectives and rules set by management are consistent with the board's overall directions. This results in a coordinated multi-level approach to control the different modes of failure.[16]

2.2.2 Flow of Authority with AI Decisions

When AI decision-making is introduced, important changes occur (depicted on top of page 7). The front-line team is now replaced by two "teams": a team of AI developers tasked to encode into machines the objectives articulated by management, together with available data, and a "team of machines" (front-line automated systems) that learns precise rules from the data for making decisions affecting customers so as to maximise the objectives.[17] The front-line automated systems both generate and execute the rules.[18]

This new decision-making regime changes and challenges the traditional governance approach as follows:

Control Gaps

We use the term "Control" to refer to management powers of delegation and instruction. By its very nature, control occurs mainly within the first line of defence of an organisation. This is a relatively familiar and well-understood task in human systems, but is challenged in AI settings – control gaps can easily emerge because of how AI works. For example:

- Management may not be used to specifying objectives with the level of precision required for AI, potentially undermining their important control task of setting clear instructions. This may lead to unexpected outcomes, since unlike humans, AI systems will not act with "common sense" when objectives are unclear or absent.
- Management may be ill-suited to delegate to AI developers due to possibly insufficient technical understanding of AI to effectively instruct them. This is exacerbated in the context of AI when compared to many other technical domains, since overall goal-setting (defining objectives) is a crucial part of AI development as discussed in 2.1.

[16] This does not mean the resulting decision making system is free of problems. There is ample evidence that human decision systems often exhibit significant bias (*Kahneman, D. (2011). Thinking, fast and slow. New York :Farrar, Straus and Giroux*) and noise (*Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: a flaw in human judgment. First edition. New York: Little, Brown Spark*). Unless appropriate mechanisms are set up to measure and audit issues with human decision making, an organisation may be "in the dark" with respect to the actual quality of those decisions.

[17] Often some rules are coded explicitly by the AI developer, but with the increased adoption of machine learning a growing number of rules are now coded by the machine itself (based on a precise mathematical specification of the objective provided by the AI developer).

[18] More precisely, the front-line automated systems include both the machine learning algorithms (that generate rules) and the resulting AI models (that execute said rules).

- AI developers may be unfamiliar with the business domain,[19] and this may result in them misunderstanding instructions from management, which may include domain-specific jargon. By contrast, human teams receiving instructions will typically have a better understanding of the business domain.
- When a staff member interacts with a customer, they can be sensitive to the particular situation of that customer and adapt their behaviour according to the customer's needs. AI systems can't relate to individual customers or leverage the richness of human-to-human experience, and important nuances will be neglected.
- With an AI system, all discretion is lost when dealing with customers – and with it any element of common sense or judgement, which frontline humans use routinely as key control measures in traditional systems.

Monitoring Gaps

Monitoring refers to mechanisms to give management and the board awareness of the activities of their subordinates, in order that they can appropriately exercise control. As such, it is a core component of the first line of defence of an organisation. Second and particularly third lines can assist in valuable ways by providing a degree of independence in monitoring, ensuring issues are properly surfaced. Monitoring mechanisms across all lines of defence can break down when AI is introduced. For example:

- The rules of customer engagement, now machine-generated, may be unreadable and hence hard for management, auditors and other stakeholders to understand or challenge.
- Front-line automated systems may make decisions without adequate transparency. Auditors who are used to asking humans why they made a decision may find it challenging to interrogate an automated system, particularly if it does not produce records of the reasons for its decisions that an auditor can understand.
- The highly technical nature of the work of AI developers may lead to a communications gap between them and more generalist management staff. If management cannot understand an AI developer, their monitoring will likely be ineffective.
- Humans monitor their own decisions when interacting with customers, and in so doing leverage their natural ability to understand context and nuance and so respond suitably – where the organisation's rules allow this response to occur. Machines are

[19] When managers delegate to a front-line team, they are often delegating to people with direct and ongoing experience of the issues at the front-line of the domain (whether that is insurance, banking, health or other). On the contrary, when management delegates to AI developers they are delegating to generalists removed from dealing with customers and front-line issues.

incapable of doing the same and cannot self-assess how reasonable their own interaction is with customers.

- AI systems don't see a person but a datapoint. The data representing a customer is a very crude approximation of the person behind it, to say the least. This directly translates to a gap in monitoring.

Next we look into how these emerging gaps challenge best-practice corporate governance.

2.3 Impacts on Corporate Governance

According to the definition from the Australian Institute of Company Directors (AICD), corporate governance "refers to the framework of rules, systems and processes put in place to control and monitor an organisation." The gaps noted above, then, threaten these two key functions of governance.[20]

Below we *transcribe* the "guiding principles of good governance" from the AICD, while **interleaving** questions that may be relevant for Boards to consider in view of these guiding principles when using AI for decision making:

1. *The board plays a key role in approving the vision, purpose and strategies of the organisation. It is accountable to the organisation's members as a whole and must act in the best interests of the organisation.*

   Machines are performing functions in the organisation and interacting with customers and clients. Does the board know if their actions are consistent with the board's stated vision, purpose and strategy, and thus "in the best interests of the organisation"?

2. *The board sets the cultural and ethical tone for the organisation.*

   Is the ethical tone set by the board translated accurately to the decisions made by the machines? Are the algorithms designed with this ethical tone in mind?

3. *All directors should exercise independent judgement and provide independent oversight of management.*

[20] *Guiding Principles of Good Governance.* Australian Institute of Company Directors. http://www.companydirectors.com.au/~/media/resources/director-resource-centre/governance-and-director-issues/guiding-principles-of-good-corporate-governance.ashx?la=en

Management delegates decisions to AI developers, who delegate decisions to AI. AI developers may only have a narrow view of how the AI impacts customers, and may not be familiar with the business domain. Are directors appropriately overseeing how the authority delegated to management is ultimately exercised?

4. *Taking into consideration the scale and nature of the organisation's activities, the board should comprise an appropriate number of directors who have a relevant and diverse range of skills, expertise, experience and background and who are able to effectively understand the issues arising in the organisation. Where practicable, the chair of the board should be independent, with the role of the chairman being separate from the role of the CEO.*

   Is there among board members an adequate understanding of AI, how it works, what it can do, how it's used in the organisation, and what are the known and foreseeable risks arising from its use? Is this understanding sufficient for them to responsibly perform their role, and comparable to the board's understanding of other technical subjects?

5. *The board should have an appropriate system of risk oversight and internal controls put in place.*

   Does the board have an appropriate framework to address known and reasonably foreseeable risks involved in using AI systems for decision making - and an ongoing means of assessing emerging risks? Are internal controls suitably tailored to the unfamiliar risks posed by AI systems?

6. *Directors should act diligently on an appropriately informed basis and have access to accurate, relevant and timely information.*

   Are directors well-informed about all relevant aspects of how AI systems are impacting people? Can the AI systems even provide such information in a manner that a director would understand? Do management regularly inform directors about how AI systems are affecting people, and are they held to account if information is not readily provided?

7. *The board would normally delegate certain functions to management. Where it does so, there should be a clear statement and understanding as to the functions that have been delegated.*

   Functions delegated to management are ultimately performed by machines. Can the board ensure that the instructions provided to

machines correspond to the intentions with which they were provided? How would the board know if there was a gap between its intentions, and the decisions of the automated system?

8. *The board is responsible for the appointment of the CEO and the continuing evaluation of his or her performance.*

   Is the board appropriately resourced to evaluate the performance of the CEO against management of the AI risks that the company incurs?

9. *The board should ensure that the organisation communicates with members and other stakeholders in a regular and timely manner, to the extent that the board thinks is in the best interests of the organisation, so that they have sufficient information to make appropriately informed decisions regarding the organisation.*

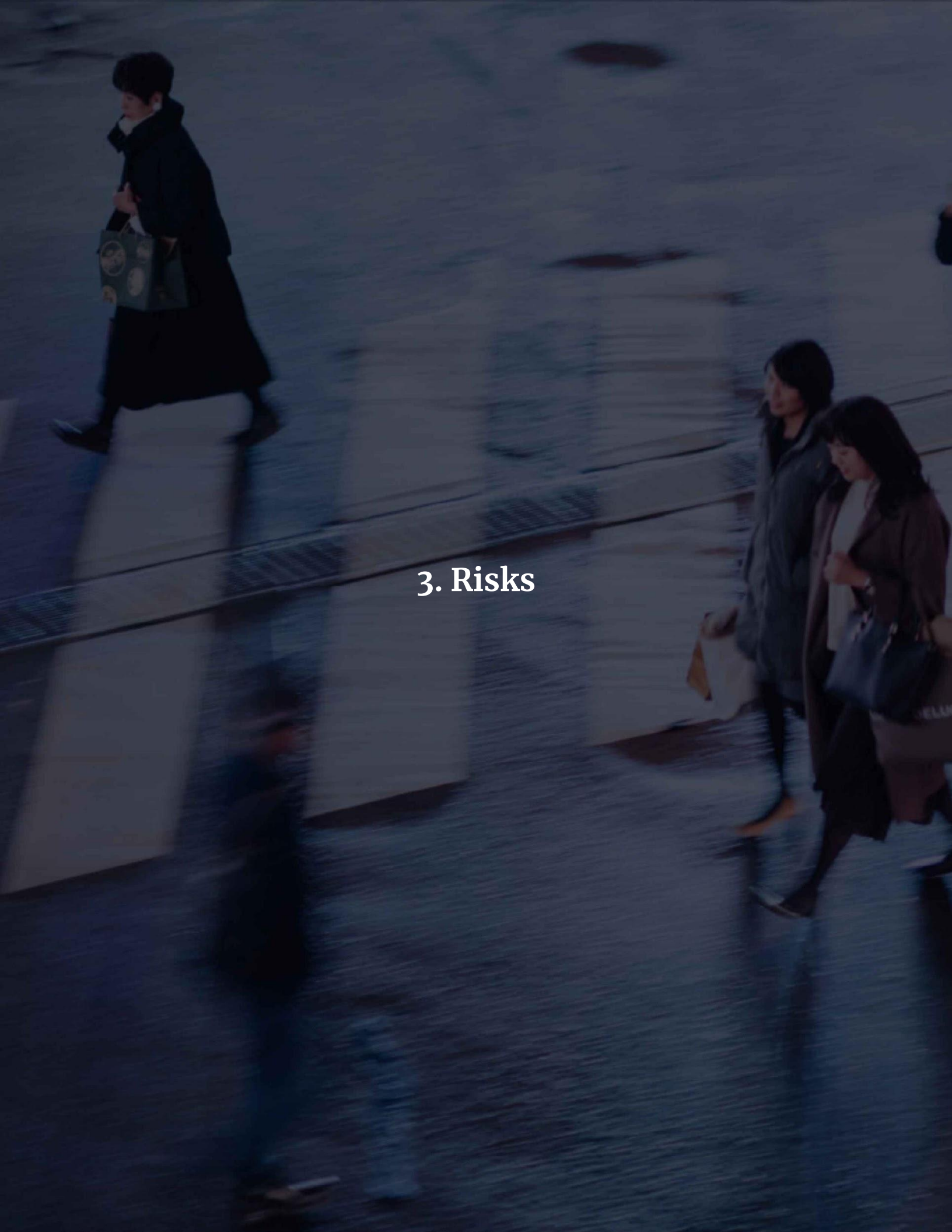
   Can the board ensure that the views of stakeholders are incorporated into the decisions being made by the organisation's AI systems?

10. *The board's performance (including the performance of its chair, the individual directors and, where appropriate, the board's subcommittees), needs to be regularly assessed and appropriate actions taken to address any issues identified.*

    Is the performance of the board and its members appropriately assessed against their ability to ensure appropriate control and monitoring of the use of AI decision making? Is the board appropriately engaged with the question of AI decision-making in the organisation?

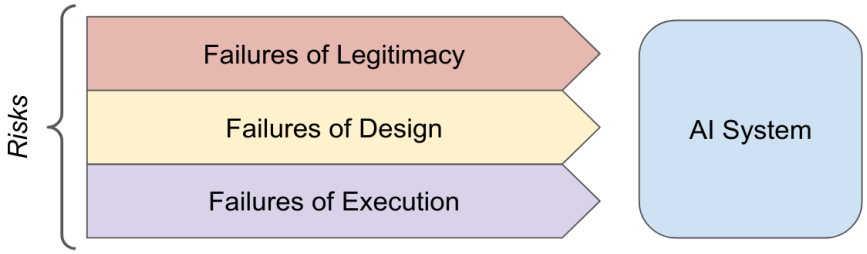
Clearly, the introduction of AI systems into organisations for use as part of governance infrastructure presents an emerging area of risk. Boards and senior management teams should be concerned, and take steps to address emerging control and monitoring gaps. While the list above highlights some questions a board ought to be able to answer, management must also take responsible actions to ensure good governance at their level. We highlight various activities that management might consider to help reduce the control and monitoring gaps in Section 4.

Before doing so, we provide in the next section a more detailed analysis of the gaps by describing specific types of risks they can manifest, with exploration of each via a series of general examples. This will assist us in determining appropriate actions to narrow the gaps.



3. Risks

In this section we identify, explain and illustrate a range of risks associated with the use of automated decision systems. We will also illustrate how each of these risks is associated with the overall loss of traditional models of control or monitoring.



As illustrated in the diagram above, we put these risks into three broad categories: failures of legitimacy, design and execution. (Appendix C provides a more detailed diagram.)

3.1 Legitimacy

Legitimacy is about general acceptance that an authority has the right to exercise power within a certain scope. If a company or government uses its powers illegally, or in ways that customers or the broader society believe is unacceptable, it faces the risk of being sued, losing customers or harming its reputation. In such cases there is a lack of legitimacy.

Below, we illustrate a few themes of how the emergence of control and monitoring gaps can lead to failures of legitimacy.

3.1.1 Unlawful Process

The use of AI systems to make decisions previously made by people can result in decisions that are unlawful. In some cases, the law could require a human to exercise discretion rather than automate a task, or the process of automating a task could involve simplifying or standardising decisions in a way that is contrary to legal requirements for a nuanced, fair, or reasonable decision. Or an automated AI system could incorporate illegal factors into a decision (contrary to, for example, anti-discrimination law). In human systems, we rely on human beings knowing or being trained on the requirements of the law and avoiding breaking the law; in many cases an unlawful act by a subordinate should be detected by regular monitoring activity, while an unlawful instruction by a manager should be identified

by an effective risk or audit function, or a subordinate blowing the whistle. But as we discussed previously, both control and monitoring are undermined by AI, potentially reducing the effectiveness of traditional governance measures. An AI system does not know what the law is and will not observe it unless legal requirements are built in as objectives or constraints; mistranslation of board and manager directives (discussed further below) can result in illegal actions.

In 2015, the Australian federal government deployed an automated system to recover welfare overpayments based on data-matching between tax office data and welfare data. Previously, data-matching was used to identify potential discrepancies, a subset of which would be reviewed by a human who would gather more material to establish to their satisfaction that there had been overpayment, as the law required. The newly introduced system used data-matching to trigger a letter to welfare recipients; if recipients failed to respond a debt was raised against them. This was unlawful because Centrelink could not be satisfied, based only on data-matching, that there was a debt: assumptions made in the data-matching, in particular that income reported to the tax office was received evenly over the year, were known to be often wrong.[21]

It seems that the goal of automating recovery of overpayment, in the absence of accurate data on when and how income was earned, led to mis-specifying the problem – a more general problem described below at 3.2.2. This led to widespread errors when sending debt notices automatically, causing distress, financial hardship and mental health problems for thousands of people. This was particularly common and impactful for already disadvantaged individuals with irregular sources of income.

This "robodebt" scandal resulted in multiple enquiries by the Attorney General and Senate. Ultimately, class action led to the Government agreeing to declarations by the Court that debts raised purely as a result of data-matching were not valid. The Government agreed to repay over $750 million dollars unlawfully recovered by the system, pay $122 million dollars in compensation, and wipe out more than $1.7 billion in outstanding unlawful debts raised by the system.[22]

[21] *Prygodicz v Commonwealth of Australia (No 2)* [2021] FCA 634; see also the discussion in Terry Carney, 'The New Digital Future for Welfare: Debts Without Legal Proofs or Moral Authority?' (2018) *UNSW Law Journal Forum* 1-16, https://www.unswlawjournal.unsw.edu.au/forum_article/new-digital-future-welfare-debts-without-proofs-authority/.

[22] *Prygodicz v Commonwealth of Australia (No 2)* [2021] FCA 634.

There are many areas of the law that are directly relevant to automated decision making, such as privacy, anti-discrimination and consumer law, among others. When deploying an AI system, an organisation (be it a company or a government department) needs to be just as confident that it's not breaking the law as when not using AI. It may need to adapt traditional monitoring and control processes to ensure it can be confident of the legality of its actions.

3.1.2 Lack of Social Licence

Even if a certain use of technology is entirely legal, sometimes organisations have strong incentives to avoid using it because of a lack of public acceptance and a concern for reputation. However, control and monitoring gaps driven by automated systems may mean that socially unacceptable systems are still deployed. Not only may there be a lack of understanding across the organisation of the likely impact of a system, but also it may be harder for managers and the board to obtain feedback about its failures. In human systems, socially unacceptable outcomes are in many cases met with immediate feedback from those affected to humans in the front-line. Those monitoring the activity – if effective in that role – will quickly understand this and can act to correct it. AI systems may not be as effective at receiving such feedback, and those affected may be less able to or willing to give it, for instance due to concerns that the system will be unlikely to change in the face of the feedback.

In 2018, Buolamwini and Gebru published a paper finding that the accuracy of the IBM and Microsoft face recognition systems were significantly higher for white people and for men than for women, people of colour, and especially women of colour, for whom accuracy was the most compromised.[23] Public outcry ensued, and in 2019 researchers reassessed the performance of the systems to see if there had been any change.[24] The authors found that the performance had improved overall and the gaps reduced, although gaps remained. In this second study, the authors also audited the performance of Amazon's system – which wasn't exposed in the first study – and found that the overall performance was inferior and the gaps much larger (Amazon contested these findings[25]). This led the researchers to conclude that public pressure had been more effective at changing the behaviour of the exposed companies. The study concluded that all companies surveyed still had issues with inequitable performance across gender and race.

[23] http://gendershades.org/overview.html. Both Microsoft (http://gendershades.org/docs/msft.pdf)and IBM (http://gendershades.org/docs/ibm.pdf) responded at the time of the paper that they already had more accurate systems under production, shortly to be released.

[24] Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. I.D. Raji and J. Buolamwini, AIES 2019.

[25] https://aws.amazon.com/blogs/machine-learning/thoughts-on-recent-research-paper-and-associated-article-on-amazon-rekognition/

Automated facial recognition systems for surveillance are often employed differently from their manual counterparts, in a manner that jeopardises their social licence. Human operators searching for suspects are incapable of trawling through thousands of hours of footage from hundreds of cameras. As a result, they must use their knowledge of the suspect to refine the search to locations and times in which they believe the suspect may be present. A potential facial match is supported by other evidence suggesting that the suspect would be in that location. On the other hand, AI systems make it cheaper and easier for investigators to cast a wide net and produce potential matches across a city-wide network of cameras within minutes. Even a very accurate system will still produce many errors given the large number of candidate faces that it can process, resulting in false arrests and a degradation of trust in the system.

Following ongoing public pressure on the use of face recognition technology, IBM eventually announced in 2020 that it would no longer offer face recognition products. It was soon followed by Amazon who instituted a one-year moratorium on the use of their technology for policing. Microsoft was next saying it would not sell face recognition to the police until the government passed federal legislation regulating the technology. In November 2021, Facebook announced it was "shutting down its face recognition system".[26]

3.1.3 Lack of Transparency

Some systems must be opaque (to certain parties) in order to operate effectively. For example, software to detect money laundering would be ineffective if it was entirely visible to money launderers. However, in many circumstances, transparency of some form may be required in order for the decision to be legitimate: for example, sometimes reasons must be given for a decision. In human systems, this can often be resolved by way of conversation, for example with a service agent. Automated systems require substantially different mechanisms to ensure adequate transparency, particularly in order for monitoring to occur effectively. The case study in the box below illustrates how issues with lack of transparency can undermine legitimacy.

[26] https://about.fb.com/news/2021/11/update-on-use-of-face-recognition/

NarxCare is a US prescription drug monitoring program: a set of databases and algorithms that track patients' use of potentially addictive drugs. Hospitals and pharmacies use the algorithm to search for patients who are concealing an opioid addiction by sourcing pain management drugs from multiple doctors. For such patients, the system can alert doctors, who may direct them into a support program, or stop issuing the drugs. But for the majority of people presenting with legitimate pain, an erroneous flag may be excruciating if it leads to the doctor refusing to prescribe medication.

The company that makes NarxCare, Apriss, uses a proprietary model and dataset. The algorithms used, and the features used to train these models are not clearly disclosed to the public or to doctors. In a media report about the system, patients affected by the system's flagging procedure complained of ambiguous or contradictory information about the extent to which the location of medical appointments, or past medical diagnoses were used to inform the decision.[27] According to this media report, a patient's risk score may even be affected if they buy medication for a chronically ill pet.

Where datasets and models are proprietary, doctors may use a system without knowing what its data or training objectives were. For example, the previously mentioned media report referred to a study in which 20% of the system's erroneous flags were on cancer patients, who would of course require visits to multiple medical specialists, but neither the patients nor doctors knew whether the system was compensating for this. As such doctors did not have access to the evidence to evaluate whether they should or shouldn't accept the system's recommendation. This lack of knowledge of the operations of the system by the doctors is self-evidently a monitoring gap.

When decisions are made only by people, doctors can seek an explanation as to the reasons for the recommendation and make an informed decision based on that evidence. This may not be possible with a proprietary AI system. A monitoring gap may be manifested, and can undermine legitimacy of a system.

[27] The pain was unbearable. So why did doctors turn her away? M. Szalavitz, 2021, Wired https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/

3.2 Design

"When you pass from the vague to the precise... you always run a certain risk of error." [28]

A system's goals may be legitimate and yet they may be poorly translated into a quantifiable metric that the AI takes actions to optimise. This introduces a control gap as the system's objectives are no longer necessarily aligned with the intent behind its creation, potentially resulting in unintended outcomes.

Although all systems are subjected to a risk induced by the inaccurate design and encoding of general intent into metrics that can be operationalised, systems that rely exclusively on human decisions have greater flexibility and feedback mechanisms that can reduce the impact of misspecified objectives. Simply put, humans may decide that a standard process is "wrong" and act unilaterally to "correct" it. While this human agency can be a source of problems, it does help to ensure rules are not applied so rigidly as to cause issues. AI systems, on the other hand, have no in-built moral compass, no understanding of societal norms nor any concern for potential repercussions as a result of its decisions outside of what they have been explicitly instructed to optimise. The computer will do exactly as it has been told – even if there would be broad agreement among people that the decision is "wrong".

In this section, we will review some of the common factors that contribute to the poor design of the intended objective into an AI System.

3.2.1 Missing Normative Constraints

When humans make decisions, they are unaware of many "common sense" constraints and rules they apply. As such, when people specify to a machine what needs to be done they will tend to miss many things that are important without realising. This is a common and often serious control gap.

For example, a marketing campaign for groceries set to maximise sales may determine that the optimal way to achieve this goal is to focus entirely on marketing of unhealthy foods. The marketing manager likely didn't intend this outcome when instructing the AI developers to "maximise sales". In this case, a limit on the amount of unhealthy foods which can be marketed may need to be imposed so as to avoid this unintended consequence.

[28] Bertrand Russell, The Philosophy of Logical Atomism, Lecture 1 1918-19.

In 2015 a misclassification error by Google's photo app applied the racist label "gorillas" to two black subjects, in line with historical stereotypes that dehumanise black people in Western society.[29] All classification algorithms make mistakes, but when designing them we can try to choose which mistakes we want to avoid most. Mistaking a chair for a table is unlikely to have moral significance. But dehumanising people of colour is profoundly damaging. Humans (should) know this and avoid such damage without being told, but machines need to be instructed. Designers can consider the social and historical circumstances of their systems and apply rules to account for them. In this case, that may have been the conservative decision to omit labels that correspond in any way with stereotyped depictions of racial minority groups. Still, doing so for all potentially harmful miscategorisations could prove to be a formidable challenge.

To design a complete set of rules to adequately constrain the behaviour of an AI system is beyond current capability – that would amount to a “holy grail” of integrating human values into machines. Nevertheless it’s imperative to keep pushing in that direction and dedicate best efforts to identify critical constraints that can prevent a system from causing harm.

3.2.2 Poor Proxies

Sometimes we don’t have the data we would like to have. It may be hard to directly obtain access to measurements corresponding to the true quantity of interest, and a proxy has to be used. If there is lack of sufficient care when picking a proxy measure, issues of design can arise, representing a control gap. The case study presented in the box below illustrates this situation.

[29] https://www.bbc.com/news/technology-33347866

One study assessing a proprietary algorithm that estimated a patient's health risk for the purposes of directing them to a special care program highlights this issue well.[30] In the absence of detailed biomarkers, the designers of the algorithm used individuals' healthcare costs as a proxy for their healthcare needs when training an algorithm to predict healthcare needs for future patients. This is perhaps intuitive, since sicker patients tend to incur greater healthcare costs. However, the study revealed that, for a given healthcare cost, black patients had a higher true health risk than white patients (as measured by biomarkers the researchers subsequently collected). Since the algorithm made triaging decisions based on cost, this meant black patients were denied special care at a higher rate than white patients for the same level of true health risk. Further investigation pointed to a key factor explaining why cost of service was a biased proxy: black patients on average had poorer access to health services than white patients, so the amount of investment in the health of black patients was on average lower for the same level of health risk.

In conclusion, the designers expected cost of service to be an accurate proxy of health and although it is indeed correlated with health, it proved to be a biased proxy, and the outcome perversely affected the disadvantaged group.[31]

3.2.3 False Causality

In some cases, an AI system may be designed using correlational information, but instructed to act in a manner assuming the information is causal. This is a subtle but potentially devastating control gap. If the AI's actions change the properties of the environment under which its training data was collected, the correlations that it identified may no longer hold, producing wildly inaccurate predictions.

[30] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447-453. doi: 10.1126/science.aax2342. PMID: 31649194.
[31] According to a later report in *Nature* (https://www.nature.com/articles/d41586-019-03228-6), the company which developed the algorithm repeated the researchers' analysis and found the same results, although the company also in a response noted that it considered the researchers' conclusion to be "misleading", as the cost model was not intended to be a sole determinant selecting patients for clinical engagement. *Nature* reported that Obermeyer was working with the company to improve the algorithm.

An AI system was developed in the mid 90's to identify patients that were at high risk of dying from pneumonia upon hospital admission, based on historical case outcomes and information such as the patients' symptoms and their medical history.[32] The idea was to use the predictions to inform whether the patient should be admitted or sent home.

When the data scientists inspected the results of the system, they discovered that if a patient had a history of asthma attacks the system was more likely to predict low risk. This struck the data scientists as odd because asthma is a potentially fatal pre-existing condition when a patient presents with pneumonia. When the doctors were informed of the finding, they recognised the reason: asthmatic patients were immediately admitted to hospital and closely monitored by staff resulting in an even lower death rate than for the patients sent home.

The system should have been designed to explicitly answer the question "what *should be done* with a patient presenting with pneumonia?". Instead it was designed in a way that unintentionally answered the question "what *happens* to a patient presenting with pneumonia?" The design was correlational when it should have been causal. This was a critical control gap that fortunately didn't translate into a disastrous decision due to appropriate human oversight early in the process.

Were decisions in the triaging system purely human, only humanly understandable criteria would have been used to make the decision of how to triage the patients. These would have been based on known causal factors, reducing the chances of triaging based on spurious correlations.

3.3 Execution

Even with a legitimate objective, and adequate design of a system to deliver on that objective, execution failure can still occur. The gaps in control and monitoring driven by AI identified previously make operational failures more challenging to address than in traditional human decision domains. This potential for operational failure is regularly identified in existing guidance on 'Model Risk Management', for example SR 11-17 issued by the US federal reserve.[33]

[32] Richard Caruana. 2019. Friends Don't Let Friends Deploy Black-Box Models: The Importance of Intelligibility in Machine Learning. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). Association for Computing Machinery, New York, NY, USA, 3174. DOI:https://doi.org/10.1145/3292500.3340414

[33] https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf

Organisations must be aware of the novel forms of execution failure associated with automated systems, compared to traditional human systems, and take steps to manage them. Some examples follow.

3.3.1 General Failure of Monitoring

Modern organisations are accustomed to continuous monitoring of critical IT systems. Monitoring mechanisms for traditional human decision systems, however, tend to be less strict. Commonly a monitoring function exists, but this may operate with a time delay, or review only a small sample of interactions. The inherent subjectivity of some human decisions also makes such monitoring less "black and white" than the performance monitoring of critical IT infrastructure.

When decisions are automated, monitoring systems need to become more IT-like. Notably, just like IT systems, automated decision making systems may fail catastrophically, and quickly. High frequency trading systems, for example, have been regularly implicated in "flash crashes".

"Tay", a Twitter chatbot designed by Microsoft, was pulled after it quickly evolved from low-stakes interactions to racism and abuse.[34] A monitoring system aware of, and looking for the potential for such outcomes may have allowed the plug to be pulled sooner.

Automated systems may also degrade in performance over time, and this degradation may be uneven. It is important to monitor system performance both as a whole and for specific groups, notably vulnerable groups, in order to spot important degradation issues as they occur.

3.3.2 Ambiguous Accountability

An AI system can have certain decisions delegated to it, but it certainly cannot be held accountable in the same manner as a human decision maker. It cannot be disciplined or fired for making poor decisions (it might be turned off, but this does not involve the system taking responsibility, or feeling consequences).

Who, then, is ultimately accountable when an AI system "goes wrong"? It should concern organisational leaders that this is a question regularly posed today. If this is an open question in an organisation, there remains a risk that *nobody* considers themselves accountable. This is

[34] https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

both a control and monitoring failure. Issues can easily fester, becoming ever more serious, when nobody is considered accountable for listening to them or fixing them.

The primary gap, in our view, is likely to emerge between junior-mid management and the developers of AI systems (which may be an internal team, or could be external for a procured system). Management may feel ill equipped to take accountability for a system they do not have the expertise to understand, and may attempt to push accountability onto AI developers. However, most AI developers are unlikely to believe they ought to be accountable for the actions of a system in a business domain where they may have little to no expertise, and where the goals of the system were determined by management. This mirrors the historical tension upon the introduction of IT systems, with traditional management being (at the time) ill equipped to understand or govern a new technology. Ultimately this was not a problem pushed onto IT - rather it led to the creation of executive level roles for IT, and general improvements in IT literacy within management teams and boards. Similarly, we feel that improvements in executive and board literacy will be warranted for AI, if it is to be governed responsibly.

Failures of accountability of this form can result in issues festering, escalating in seriousness, and then escalating up the hierarchy. We note that governance failures of this form will likely reflect badly on the senior management staff responsible for that area of the business.

A further complication is the likely implicit or explicit delegation or contracting of certain activities to either developers or vendors of AI systems. This may confuse management, and is particularly problematic for vendor relationships. Similarly to other forms of software, some delegation of responsibility – particularly for functionality – might be expected to occur at this point. However, unless specified otherwise, this should always be understood as delegation of a responsibility, not transference or abdication. The manager delegating responsibility for constructing an AI system should not expect to transfer accountability for that system, particularly for its failings. There is a risk, though, that this is misunderstood by them. For external partners, contracts may outline various responsibilities and accountabilities of each party, but it is likely that people affected will consider the original firm responsible for decisions made by an AI system. Again, there is a risk that management does not properly understand or accept that whatever might be true of the *legal* risk, accountability in the eyes of the public and customers – and the potential for reputational damage – is not transferred, which may lead to monitoring and control failures.

The recent (Australian) Royal Commission into Misconduct in the Banking, Superannuation and Financial Services Industry provides a series of case studies in consumer lending that, at least in part, illustrates problems of ambiguous accountability arising from automated systems. In at least two case studies,[35] errors in automated systems were found to persist for a significant period. These errors were impacting customers, without an accountable person either noticing at all, or taking adequate corrective action. It may be that equivalent errors within a purely human system would have been detected earlier, or at least without causing errors in relation to so many people and accounts, and corrected more quickly and effectively due to clearer accountability and the ability, when directing human decision-makers, to more easily direct them to change their approach. The Royal Commission specifically noted the absence of – and need for – 'end-to-end' accountability for the design, implementation, marketing, and monitoring of financial products.[36]

3.3.3 Error Replication

In traditional systems, a front-line team consisting of human operators limits the speed at which decisions are made and acts as a final sanity check on the overall system's decisions. A single human operator making regular mistakes has limited power, which acts to reduce the potential impact of their errors on a population. This control is removed in AI systems. A single error that gets introduced to the decision-making process could instantly impact the organisation's entire customer-base. The potential volume of affected individuals means that even minor errors may pose a significant reputational or financial risk to the organisation. This intensifies the need for effective control and monitoring of AI systems.

In January 2016, Google pushed a software update to its Nest Learning Thermostats. It was reported that an error in the code caused many of the devices to drain their batteries leaving users unable to turn on their heating in the depths of the Northern Hemisphere winter.[37]

[35] Royal Commission into Misconduct in the Banking, Superannuation and Financial Services Industry, Interim Report, Volume 2 (Commonwealth of Australia, 2019) https://www.royalcommission.gov.au/system/files/2020-09/volume-2.pdf pp 64-68, pp 74-82

[36] See in particular Recommendation 1.17 of the Royal Commission, and the discussion of accountability and that recommendation at *Royal Commission into Misconduct in the Banking, Superannuation and Financial Services Industry, Final Report, Volume 1* (Commonwealth of Australia, 2019) https://www.royalcommission.gov.au/system/files/2020-09/fsrc-volume-1-final-report.pdf , 112-116.

[37] Nest Thermostat leaves users in the cold. https://www.computerworld.com/article/3412197/top-software-failures-in-recent-history.html#slide12

3.3.4 Domain Creep

It is not uncommon for models and systems developed for particular domains, or particular use cases, to be applied in other settings which feel "similar enough". This is often described as "domain creep".

For human decision systems, the duplication of an existing function brings new front-line staff and a new rule book, and people will consult and interact with each other and use their judgement in order to identify where changes need to be made to adapt to the new domain. However, no similar controls are in place when we transfer a model or AI system to another domain.

Some forms of domain creep may be subtle, and challenging for non-expert management to spot. Perhaps more worryingly, management enthusiasm for "rolling out" an effective but narrowly tested system can be a primary cause of this form of control failure.

> During the Covid-19 crisis, many researchers attempted to build models to assist with diagnosis and prognosis. Commonly, training data for such systems was taken from a particular hospital, setting, or form of equipment. However, there was clearly a human desire to extrapolate the use of such predictions outside of this original domain – to other hospitals, settings, equipment types or countries. In several studies, it was unclear if models could be expected to perform adequately outside of the narrow domain in which the data was collected.[38]

Ethical failures can arise from domain creep of this form. A model's performance is likely to deteriorate when the model is applied in unfamiliar territory. Biases may also emerge – perhaps models perform well in a new domain, but poorly on a newly observed group of people with which the system is generally unfamiliar.

3.3.5 Input Errors

Identifying input errors in automated systems is critical to detecting failures. Human intuition will naturally correct obviously erroneous inputs, but automated systems require specific steps to achieve similar levels of correction, another manifestation of a control gap.

For example, a bank manager approving loans will likely question the applicant if they declare an income which seems unreasonably high or low. The applicant can then correct the input

[38] Roberts, M., Driggs, D., Thorpe, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 3, 199–217 (2021). https://doi.org/10.1038/s42256-021-00307-0

before it is used for a decision. An automated system will need to be specifically instructed to perform these sorts of checks. Without such checks, loans could be automatically given to those unable to repay them, or automatically refused to others without good reasons.

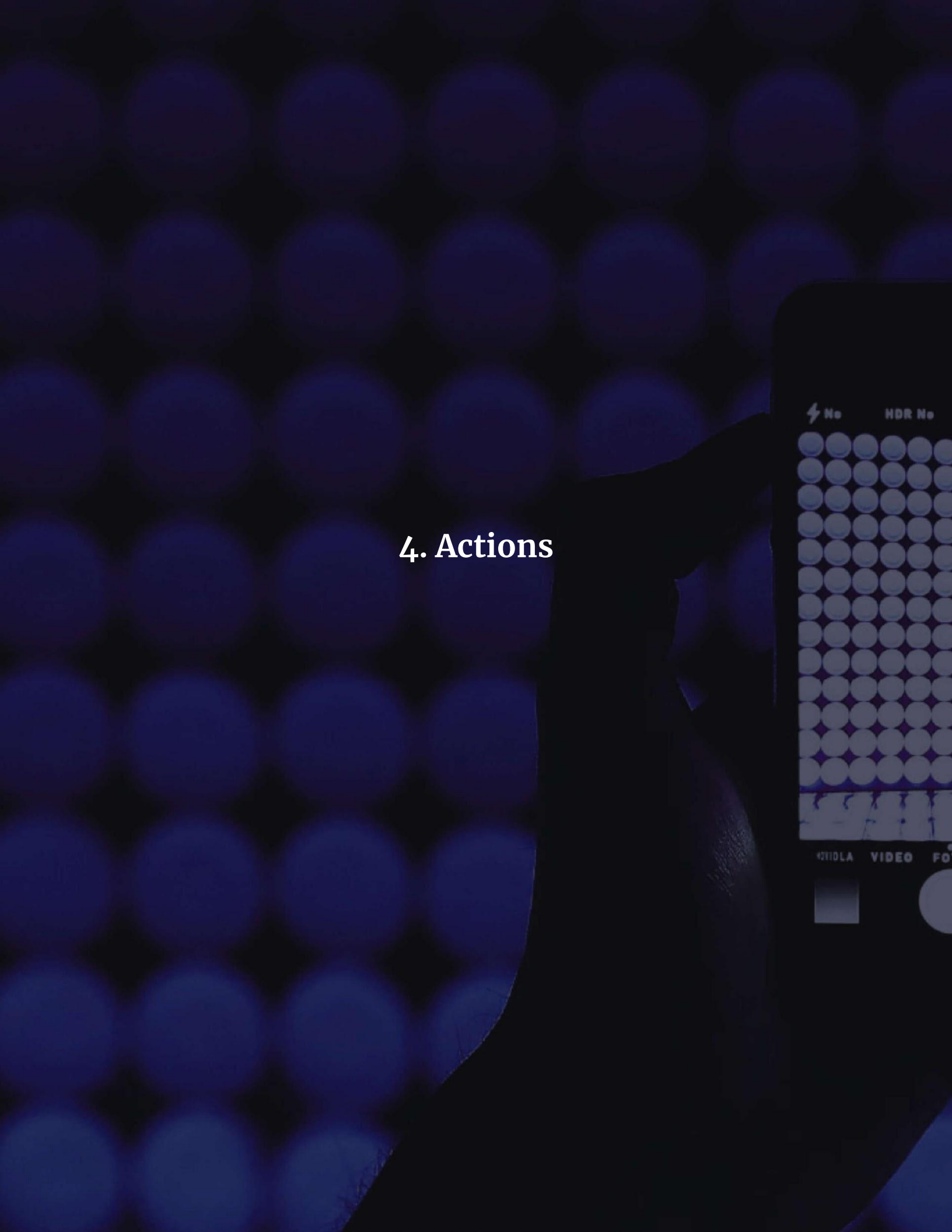
3.3.6 Off-the-Shelving

Off-the-shelving means unduly trusting outsourced expertise. This could be a client trusting, without due diligence, that a vendor has taken appropriate care in the design of an AI tool under procurement and that the tool addresses the client's needs.

For example, a police department may determine its policing resource allocation based on a purchased AI system that claims to predict crime. The system could be using past arrests to predict future crime – as if arrests were a true proxy for the real world incidence of crime, when in fact it is biased by historical patterns of arrest. If the department acts on the predictions without adequately understanding the assumptions that underpin the system, it may create a feedback loop that reinforces patterns of historical discrimination: This could result in the police only making arrests in communities that the system sends them to.[39]

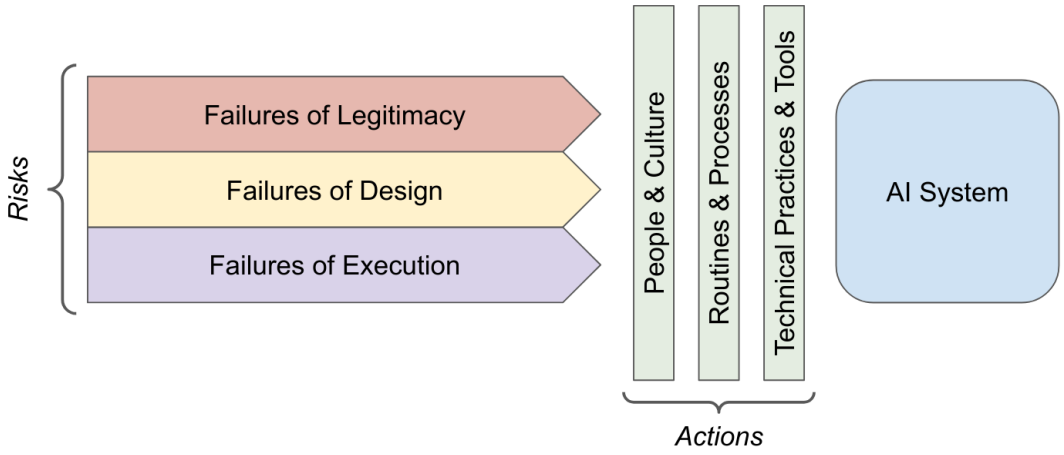
Diffusion of responsibility is another hazard associated with off-the-shelving: a naive customer is likely to feel somebody else is now responsible because they paid money for the tool. In summary, the deployment of a tool procured from outside an organisation doesn't provide a licence to ignore the need to evaluate the consequences of its use.

[39] https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained



4. Actions

As illustrated in the previous section, there are a range of serious risks to organisational governance driven by the introduction of AI systems. We believe it is possible to effectively address those risks. In this section, we identify a series of suggested actions under three broad themes: people and organisational culture, organisational routines and processes, and technical practices and tools (see figure below, and a more detailed version in Appendix D). We view these actions as elements of an emerging **responsible AI**, or **ethical AI** practice which can lead to general improvements in the governance of automated decisions by addressing failures of legitimacy, design and execution.



We encourage the reader to take these as suggestions, to be suitably adapted and augmented to meet the specific needs of their own organisation.

4.1 People and Culture

Traditional organisational governance is centred on people, and governance of AI should be no different. Fundamentally, people will be held responsible for any failings of AI. However, as observed above, significant control and monitoring gaps are emerging within AI systems, due to inadequacies in traditional skills, motivations, governance methods and culture of people within organisations. As such, interventions focussed on people and, more broadly, organisational culture, can be effective in narrowing control and monitoring gaps.

4.1.1 Appropriate Incentives

"It is difficult to get a person to understand something when their salary depends on them not understanding it." [40]

Incentives are very powerful interventions and cannot be underestimated. Sound incentive design can have significant benefits across all areas of risk. However, the poor design of incentives can increase risk.

Including explicit objectives that attempt to incorporate elements of common sense or morality into an AI system's design (as an attempt to avoid failures of design) can compromise the business performance of the system (for example reduce profits).[41] It will be hard for system owners and managers to justify such designs unless leadership explicitly incentivise or require such "ethical objectives" or constraints for AI .

For example, a public company may declare simple, high level objectives like a target profit margin, target revenue, or customer satisfaction score. These objectives may be agreed by the board and senior management team, and promoted to shareholders. However, such metrics do not adequately describe broader non-commercial objectives such as ensuring that the AI systems don't make unfair decisions or decisions which, while legal, impose social costs. Within such an environment, a middle manager with direct responsibility for controlling and monitoring a system may find it difficult to accept a lower profit or revenue in order to promote a social or ethical goal. In the worst cases, they may be specifically incentivised (for example, by bonuses or promotion rewarding outstanding profits) to do the legal 'bare minimum', and may instruct their AI developers to pursue a narrow goal in full knowledge of the ethical failures which may ensue.[42]

We therefore recommend that organisations consider mechanisms to explicitly or implicitly incentivise the development and adherence to responsible AI practices among product owners, senior management and AI developers. This will help to foster greater legitimacy. Other stakeholders such as system developers may also become more aware of responsible AI considerations in their work if they understand them to be a key concern for system owners and senior executives.

---

[40] Sinclair, Upton. I, Candidate for Governor: And How I Got Licked (1935) (original quote edited slightly to use more inclusive language).

[41] It won't necessarily be compromised since there may be "profitability plateaux". An AI system may admit several design configurations that lead to essentially the same "maximum profit", and yet some of them may be significantly better than others from an ethical perspective. E.g. see https://ambiata.com/blog/2021-03-22-nba-for-social-good/

[42] On the challenges of incorporating ethics, specifically in a Silicon Valley context, see Metcalf, Jacob, Emanuel Moss and danah boyd, 'Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics' (2019) 86(2) *Social Research: An International Quarterly* 449.

There are many potential mechanisms to incentivise a responsible AI practice for improved governance, such as:

- characterising loss of potential profit or revenue due to implementation of fairness constraints as 'investment in fairness'
- adjusting target metrics to account for a range of objectives that capture not only commercial but also broader ethical objectives
- celebrating and sharing the stories of system owners that implement responsible AI processes
- flagging implementation of responsible AI as a critical factor in annual performance reviews for system owners
- using key performance indicators that measure customer impacts
- integrating ethical considerations alongside technical considerations in the organisation's development workflow[43]
- providing specific and meaningful incentives for ethical conduct
- removing heavily conflicted individuals from key decision making roles in the system's design
- Incentivising the inclusion of underrepresented groups in key decision making roles in the system's design

We encourage the reader to carefully consider the incentive structures in their organisation, the effects of those incentives on management's control function for AI systems, and how this may be adapted to reduce the risks of poor conduct.

4.1.2 Training

The skills required for teams to implement responsible AI draw from several disciplines including social science, political science, ethics, systems thinking and governance as well as statistics and machine learning, and include concepts only recently developed in academia. It is therefore likely that most employees of organisations are under-skilled in many areas of responsible AI – even those who are AI specialists.

We recommend that organisations provide training in multidisciplinary approaches for responsible AI to AI developers, system owners, system integrators, business leads and the board, to ensure that they have the expertise and awareness required to govern systems effectively. Greater multidisciplinary training will help to reduce both control and monitoring gaps, and will generally help to create clearer lines of communication. The table below illustrates some of the key skills needed by various roles identified elsewhere in this report.

[43] Microsoft's AI Fairness Checklist, for fairness-specific considerations: https://www.microsoft.com/en-us/research/project/ai-fairness-checklist/

Role	Skills needed
developers *data scientists and data engineers developing, deploying and validating AI systems*	• understand link between design decisions and the impacts of AI systems • communicate the implications of design decisions to non-technical audiences • quantify non-commercial, "ethical" objectives into mathematical measures, understand their limitations, and integrate them into the design of AI systems • identify unintended behaviour in AI systems, investigate causes and apply mitigation strategies • understand the ways in which AI systems can ingrain existing societal biases • understand the business domain in which they are operating
business/system owners & integrators *business leader responsible for the operation of the AI systems (either developed in-house or procured)*	• understand the novel risks introduced when AI systems replace manual systems • define both business and ethical objectives for AI systems • decide how to balance those objectives when they might compete • deploy AI transparency as appropriate for the system context • understand and communicate the implications for the system's impact of technical design decisions
system review committee *cross-disciplinary committee ensuring system owner follows responsible AI process in the design, development and deployment of AI systems*	• assess the degree to which a system's design and operation aligns with its stated objectives, and the values and priorities of the organisation • understand the limitations of the system as implemented and its potential unintended impacts • ensure the organisation maintains an effective process of responsible AI governance.
board of directors *governing body that sets management and monitoring policies*	• understand the risks introduced when AI systems replace manual systems • setting strategic direction on AI use (including when **not** to use AI) • ensure that the organisation maintains an effective process of responsible AI governance

4.1.3 Broad Stakeholder Engagement and Co-design

This intervention addresses failures of legitimacy and design in general. System development should not be done in isolation from those the system might impact. To foster greater legitimacy, a diverse range of stakeholders potentially impacted by a system may be given the opportunity to give input into its design and operation. This may utilise mechanisms of consultation, community juries[44] and other methods of involving people,[45] including more structured methods for co-design. Alternatively, less formal methods of iterative design and feedback, potentially via informal design workshops, surveys, testing or prototypes, may be effective.

Whatever approach is taken, the involvement of a suitably diverse range of stakeholders in the design of a product or project will tend to reveal blindspots of the designers, identify issues before systems are put into market, allow important feedback and monitoring themes to be discovered early in the development process, and ultimately foster greater legitimacy of the final system. Importantly, these approaches will only work if feedback from stakeholders is given effect in system design and system governance, and risks they identify are taken seriously. Consultation or co-design that has no discernible effect on outcomes will undermine, not increase legitimacy. For significant or sensitive systems, mechanisms for ongoing engagement with stakeholders could be necessary.

4.1.4 Role-Taking

Role-taking is the socio-psychological concept referring to the ability of an individual to understand others' thoughts, feelings, experiences, and perspectives.[46] A broadened sense for the impact of the AI system on people can help reduce the gap in both control and monitoring functions.

The AI sector is responsible for conceiving, designing, implementing, deploying and assessing systems that will shape the lives of diverse populations. However, most demographics, particularly disadvantaged groups, are underrepresented in this sector, which is dominated by a white male workforce.[47] As a result there is a risk that organisations miss important or even critical considerations.

---

[44] https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/
[45] See eg Ada Lovelace Institute, *Participatory Data Stewardship* (Ada Lovelace Institute, 2021) <https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/>
[46] Mead, George Herbert. 1934. Mind, Self, and Society. University of Chicago Press.
[47] https://www.weforum.org/reports/global-gender-gap-report-2021

In addition to the perhaps obvious and highly recommended intervention of increasing diversity in the sector and relevant professions, progress can also be made with existing professionals. There is evidence that people from majority groups can improve their role-taking abilities through structured collaborative activities with members of marginalised populations who are empowered as leaders within an interaction situation.[48] This opens the possibility of role-taking training as a valuable intervention for improving the legitimacy of design processes for AI systems. Simple interventions can be introduced internally to improve the team's role-taking, some exemplified below:

- Microsoft's Judgement Call[49] is a roleplaying activity that can be played during the design phase of development to cultivate stakeholder empathy by imagining the effects of the system from different perspectives.
- Tarot Cards of Tech[50] is a brainstorming exercise that presents designers with hypothetical futures for the AI system. Participants are encouraged to explore the consequences of these futures with a focus on stimulating discussion around preventing undesirable outcomes.

Because role-taking is necessarily relational, it is crucial that these activities are implemented in collaboration with stakeholders from underrepresented and marginalised groups. Not only should these stakeholders be part of the conversation, but empowered as leaders within activity sessions and recognised as experts in their lived experiences. This aligns with the values of co-design, as discussed in 4.1.3, and serves key training needs as listed in 4.1.2.

4.2 Routines and Processes

Standardisation and routine is a powerful mechanism to reduce risks. For example, it may be challenging for management to specify from scratch an effective monitoring approach each time an AI system is deployed, or to ensure the full and complete specification of an AI system to a developer when exercising control, without the aid of some standardised processes to guide them.

Standardisation and routine provide a useful default position for the governance of an organisation. It fosters a baseline level of effectiveness. Standard methods may be varied – usually enhanced – with good reason, but their existence is still required to allow an organisation to effectively maintain a suitable baseline level of control and monitoring.

[48] Love, Tony P., and Jenny L. Davis. 2021. "Racial Differences in Women's Role-Taking Accuracy: How Status Matters." Sociological Science 8: 150-169
[49] https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/judgmentcall
[50] http://tarotcardsoftech.artefactgroup.com/

4.2.1 Roles and Routines to Support Accountable Persons

As observed in Section 3, organisational accountabilities may become unclear when AI systems are deployed. It is important that management are clear in their accountabilities and confident that they have the skills and support to exercise them effectively. Improvements in accountability can mirror benefits of appropriate incentive design, since a key component of accountability is sanctionability: a negative incentive. As such, just like incentive design, roles and routines serving accountability goals can have broad impacts across all three major categories of risks we identified: legitimacy, design and execution.

In our view, primary accountability for the outcomes of an AI system should normally be assigned to the person able to delegate decision-making power to it. This would commonly be a manager within the company who would have traditionally had the power to delegate the task to a group of people.

A manager who is used to delegating to people and managing them may not be suitably skilled in the risks of AI systems to competently delegate a decision to AI developers. Organisations should not assume this competency exists in their managers. While individual managers could (and, over time and with training and experience, likely will) become competent in this new domain, in the short term organisations should seek to create roles and routines to support them.

One option is to hire an individual or team specialising in AI governance and risks. This team can create policies or routines for the organisation, offer advice to management and AI developers, and conduct independent reviews of systems.

Another complementary option is to create a dedicated governance committee for an AI system or a collection of systems. First, and most importantly, individual accountability for the outcomes of a system may be clarified by the charter of any such committee. Independent advisory or audit specialists (like the AI governance or risk specialists noted above) may formally discharge their responsibilities within such forums. Other areas of the organisation can give input into system design, as appropriate, via such forums. The explicit formalisation of ethical considerations and tradeoffs at such a committee is a useful mechanism to ensure they receive appropriate attention from management. Finally, review and escalation protocols can be formalised via the committee's charter, allowing management a clear avenue for the upwards reporting of risks or issues which may be emerging.

4.2.2 Independent Technical Audits

Independent audit is an important function in most organisations. In some situations, the importance of an independent audit is considered so great that it is required by law or regulation (published financial statements being a common example). In many cases, it is the independence of the audit process which allows a challenge to occur – even with the right intent, it is often challenging to critique one's own work, systems and processes. Independent technical audits, depending on the breadth of their scope, may be able to address a range of potential failures.

For AI systems, independent technical audits can serve three important purposes.

First, an audit can unearth technical flaws in design, construction or operation which were not apparent to the system builders or owners, and which either have the potential to cause issues, or may already be causing them. This can allow system designers to correct flaws or remediate problems before they become large.

Second, an independent technical audit can give confidence to non-technical senior management or board members that systems have an appropriate level of technical oversight. This represents a delegation of management functions similar to an audit of financial statements - management can rely on the audit process to conduct the technical checks which may be too detailed or complex for them to undertake themselves. This improves the monitoring function of senior management.

Third, an independent audit can be a mechanism for an organisation to receive advice or guidance on better practices observed in the general market, which can inspire continual improvements, and hence incremental risk-reduction over time.

Audit teams do not have to be entirely independent of the organisation itself. It may be possible for an organisation of suitable scale to employ an internal audit team, or to rotate construction and audit responsibilities within an analytics function on a project-by-project basis. With internally staffed audit functions or processes, it is important to work to manage any conflicts of interest which may emerge, particularly interpersonal relationships.

A variation on a traditional audit is the concept of "red-teaming", where an independent team is tasked with robustly challenging the status quo and attempting to "break" systems. Adversarial audits of this form are common in areas such as cybersecurity, and may also be of particular relevance for identifying flaws in the design of AI systems.

4.2.3 Buyer Beware

Procuring an AI system from a third party does not absolve the organisation from accountability for its operation. This implies that, when procuring a system, the system must satisfy the same standards that would be placed on one that is being developed internally. For example, the system's explicit objectives and how they are balanced, its limitations and failure modes must all be available for analysis and review. Systems sold as proprietary solutions ask the system owner to take accountability for the impacts of a system whose objectives and constraints they were not responsible for. This action directly addresses the off-the-shelving risk identified in Section 3.3.

Arnold et al. (2018)[51] provides a good starting point for items to seek clarification on before deciding whether to incorporate a third-party application into an AI system, such as:

- What is the intended use of the service output?
- What algorithms or techniques does this service implement?
- Which datasets was the service tested on? (Provide links to datasets that were used for testing, along with corresponding datasheets.) Do the datasets have any known biases, and are they relevant/appropriate?
- Describe the testing methodology.
- Describe the test results.
- Are you aware of possible examples of bias, ethical issues, or other safety risks as a result of using the service?
- Are the service outputs explainable and/or interpretable?
- For each dataset used by the service: Was the dataset checked for bias? What efforts were made to ensure that it is fair and representative?
- Does the service implement and perform any bias detection and remediation?
- What is the expected performance on unseen data or data with different distributions?
- Was the service checked for robustness against adversarial attacks?
- When were the models last updated?

While buyers of AI systems should take responsibility for the system's actions, suppliers should not rest easy following a sale. Suitable performance must be maintained during operation. We suggest that agreements between buyers and suppliers should include provisions around supplier actions to uncover problems, processes for reporting issues experienced by buyers or by those affected by systems, and clauses requiring promptness of corrective action when issues are discovered. Similarly to provisions found in many IT service level agreements today, this may include a range of incentives for good performance or

[51] Arnold et al. (2018). FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. arXiv preprint arXiv:1808.07261. https://arxiv.org/abs/1808.07261

penalties for poor performance of the supplier in relation to the legal and ethical risks we have discussed.

4.2.4 Preemptive Contestability

Contestability of AI decisions is a frequently discussed topic, and often seen as an important element of legitimacy. However, it is a poor outcome for all concerned if an AI system makes an incorrect decision which must then be contested, often via a formal, adversarial process. Where mistakes are expected to occur frequently, a preemptive process can be deployed where decisions can be contested before they have impacted a person. This can allow many issues to be resolved before they escalate to a full-blown dispute. It can significantly improve the social licence of a system.

One common method for achieving this in practice is to display the data relied on for the decision to the person affected, together with the preliminary decision and a clear route for the person to point out any mistakes before that decision is finalised. This directly addresses the risk of input errors identified above. There are examples of processes like this today.

A recent case involving a computer vision algorithm is instructive. A fine was issued incorrectly to a motorist, after a number plate detection system mistook text on another person's t-shirt for their car's number plate.[52] The photos of the alleged infraction were provided to the motorist, who was able to call and have the fine removed before paying it. While there are perhaps improvements to be made on this process, it is substantially superior to a fine being issued without any other information, followed by an adversarial process for a motorist to then attempt to prove their innocence, which might otherwise have occurred.

Organisations deploying AI systems should expect them to make novel, unusual and strange errors. While an organisation might try to identify mistakes before customers are impacted, it may not be possible to identify all of them. In such situations, a clear, non-adversarial mechanism for those affected by an obvious mistake to point it out and have it corrected – before the mistake significantly impacts them – is desirable, and often easy to implement. We call this "preemptive contestability".

4.2.5 Questionnaires and Checklists

A large element of risk management is the identification of potential problems which could occur. Questionnaires and checklists can help by providing a standardised starting point for reflection. Used well, they can facilitate imagination and creativity that might identify blindspots within the development team as well as cultivate a more active and engaged attitude towards the risk management process.

[52] https://www.theguardian.com/uk-news/2021/oct/18/motorist-fined-number-plate-t-shirt

Depending on how specific or broad they are, checklists and questionnaires can address few or many of the risks we identified. We encourage organisations to develop standardised checklists for use in AI development, designed to meet the scale, nature and needs of the organisation. To ensure this control is effective, organisations should consider whether completion of checklists should be mandatory prior to release of an AI system.

A number of online resources exist to facilitate the identification of potential risks during the design phase of the development process. Three in particular that we'd like to highlight are The Open Ethics Canvas,[53] MIT's AI Blindspot[54] tool and The Ethical OS Toolkit.[55] In the appendices, we have also included a pair of questionnaires that we have developed and found useful in the past, which are aimed at getting a rough sense for the risk associated with a specific system and whether an AI solution is suited to a particular application context (Appendix A and B).

4.3 Technical Practices & Tools

Interventions focused on people or processes may be ineffective without complementary technical practices or tools that are designed to suit the situation. In this section, we identify a series of technical interventions which can help to improve management's ability to exercise control and monitor AI systems. We suggest that these can complement many of the interventions already listed above.

4.3.1 Dashboards and Control Panels

The people responsible for a system must understand its impacts if they are to monitor and control it. As these impacts can change over time, managers may benefit from automated visualisation and the display of live impacts via a software dashboard. Many such tools currently exist,[56] though by default typically quantify only model performance measures like accuracy. The definition and measurement of an AI system's important impacts is a precursor to deploying any impact dashboard, and remains a context-specific challenge that must be undertaken by organisations on a per-system level.

---

[53] The Open Ethics Canvas is a tool for developers and system owners to initiate conversations with the intention of revealing blindspots in the development process https://openethics.ai/canvas/

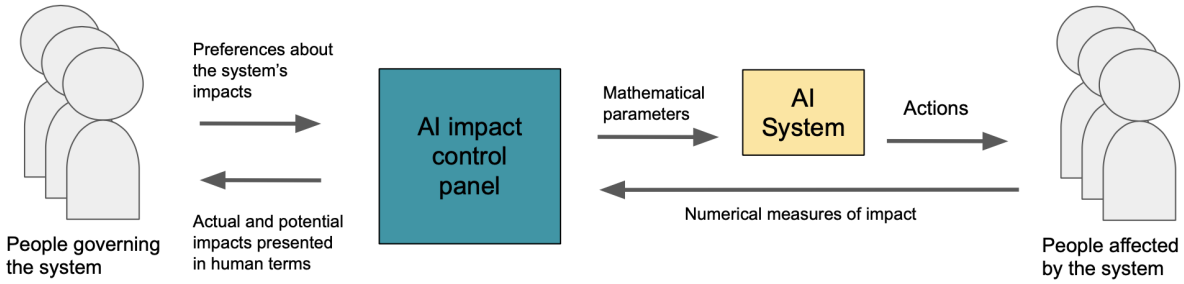
[54] MIT's "AI Blindspot", a resource to spot unconscious biases and structural inequalities in AI systems: https://aiblindspot.media.mit.edu/

[55] The Ethical OS Toolkit is an aid for organising workshops aimed at identifying risks specifically targeted at AI systems. https://ethicalos.org/

[56] Prometheus and Evidently AI are two (of many) open-source software monitoring tools: https://prometheus.io/ and https://evidentlyai.com/

However, even with an understanding of the system's current impacts, managers cannot effectively control the system without also understanding the potential impacts that could result from changing how it operates. Such lack of visibility is one of the causes of the control and monitoring gaps as described in Section 2.2.

AI impact control panels aim to close this gap by giving managers levers to control a system's impacts in an intuitive way. Like a dashboard, a control panel translates mathematical measures of performance into human-readable summaries of impact, but it also works in the other direction: translating a manager's preferences and intentions for the impact of the system into the mathematical parameters that tune the AI system accordingly. This two-way translation is depicted below.



Tools such as AI impact control panels are especially important when the managers of an AI system do not have the technical background to engage directly with the technical implementation details. A control panel enables them to examine different potential impacts the system could have, without needing to understand what technical details control those impacts. For example, a manager could examine different potential fairness / revenue trade-offs for their system and select an acceptable balance of these objectives. The control panel software then translates this choice into the value of the mathematical property that controls the system.

Like dashboards, building an impact control panel for an AI system requires system-specific input and customisation. Unlike dashboards, however, work to develop off-the-shelf software to support their implementation is at an early stage.[57]

[57] See an open-source prototype illustrating the control panel ideas: https://github.com/gradientinstitute/ai-impact-control-panel.

4.3.2 Technical Documentation and Transparent Development

The life cycle of an AI system from concept through to deployment is complex and involves numerous consequential decisions made by a large variety of contributors at different levels in the organisation. Documentation is both an artefact that brings clarity to this complexity and also a process that, if done correctly, integrates responsible thinking throughout the entire system life cycle.

To ensure that people developing, monitoring or auditing the system have a complete picture of the system to meet their needs, organisations should document all of the important design decisions, system objectives, constraints and performance indicators of an AI system in a central register of AI systems within the organisation. This should be done in a manner accessible to all internal stakeholders. At times, this may mean that documentation becomes extremely technical and detailed. While this may be a form of documentation unfamiliar to some in middle or senior management, those managers should encourage the existence of such documentation as a matter of effective monitoring. However, those managers may also request a summary of the more pertinent points for their consideration.

There are a number of resources available that propose methods for documenting the development process of the system that also place a strong emphasis on governance considerations.

- The ABOUT ML resource library[58] is the result of an ongoing collaborative effort across a diverse range of organisations within the AI community in an attempt to establish a set of best practices for documentation that industry should adhere to. It contains useful guidance and examples of transparent development that can be tailored to match the requirements of the stakeholder.

- M. Arnold et al. (2018)[59], T. Gebru et al. (2018)[60] and M. Mitchell et al. (2019)[61] provide useful templates to create accessible summaries for the AI systems, data and predictive models, respectively.

The public is another crucial stakeholder and the degree of visibility that it has over the system should be carefully weighed. Public visibility over a project can help ensure its legitimacy and is an effective method for identifying blindspots within the development process. The New Zealand Government has committed to clearly explaining to the public

[58] ABOUT ML https://partnershiponai.org/workstream/about-ml/
[59] Arnold et al. (2018). FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. arXiv preprint arXiv:1808.07261. https://arxiv.org/abs/1808.07261
[60] Gebru et al. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010. https://arxiv.org/abs/1803.09010
[61] Mitchell et al. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). ACM.

when their decisions are informed by algorithms.[62] The explanations may include information about the data used and how it is processed as well as clearly identifying a point of contact for public inquiries about the algorithm. However, transparency at a public level should always be done with consideration for the privacy of affected individuals and the potential for malicious actors to game the system.

4.3.3 Explainable AI

While people typically make predictions based on causal relationships, AI systems do so by transforming and combining data to discover complex correlations between its inputs and the output variable. Although this approach can provide impressive predictive performance, tracking the AI's internal logic can be very unintuitive to a human.

Explainable tools and models focus on answering why an AI made a certain decision. Making the mechanics of AI understandable improves both the ability to control and monitor the AI: it enables control (by ensuring the system is doing what it is meant to do) and monitoring (by alerting people at different levels of an organisation to identify problems and escalate if necessary). It can help address problems of design and execution, identifying incorrect assumptions made during its design or mistakes in programming that mean it doesn't work. It can also improve legitimacy through better transparency, as illustrated in the example in 3.1.3.

Techniques for gaining insights into the factors that influence an AI's decision vary depending on the class of model, the expertise of the user and nature of the question. Interpretable Machine Learning by Christoph Molnar[63] provides a comprehensive overview of the common approaches. The UK's Information Commissioner's Office has also published practical advice[64] to help organisations explain the decisions of their AI systems to affected individuals.

It is important to note that most explanations rely on providing a simplified view of models to make them more understandable. These simplifications make assumptions that may not hold up in certain situations. It is important that the interpreter understands the limitations of the explanation when deciding how to act.

---

[62] Algorithm Charter for Aotearoa New Zealand https://data.govt.nz/assets/data-ethics/algorithm/Algorithm-Charter-2020_Final-English-1.pdf

[63] https://christophm.github.io/interpretable-ml-book/

[64] https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence/

5. Conclusion

In this report we provided broad guidance for organisations to reduce the risks of adopting AI systems for decision-making. We described how AI systems make decisions differently from people and how those differences create both control and monitoring gaps within a system of corporate governance. We identified three broad categories of possible governance failures associated with these gaps: failures of legitimacy, design and execution. Within each of these, we recognised, explained and illustrated with real case studies a range of specific risks. Finally, we suggested actions to address those risks across multiple fronts: people and culture, routines and processes, and technical practices and tools. Details of risk management practices will vary across different organisations, and the instantiation of this guidance within any particular organisation should be in accordance with its business practices and use of AI systems for decision-making.

6. Appendices

A. AI Suitability Questionnaire

Before investing time and effort into developing an AI solution, it is crucial to ask whether an AI solution is even the right approach.

Over the last decade, advancements in AI have led to significant improvements in some automated decision-making tasks. Because of this, however, AI is often overhyped as the answer to all automation problems. In reality, modern AI systems rely heavily on a number of key assumptions and can fail catastrophically if deployed in situations where these assumptions are violated. The questionnaire below identifies several key characteristics of a problem setting that will be paramount in determining the success of an AI solution.

- **Can we quantitatively measure how well a system is satisfying its goal? (critical)**

  An AI system "learns" by scoring its current performance using some quantified metric (such as the accuracy of its predictions on sample data) and adapting its behaviour so as to increase that score. If the system's goals cannot be quantified, the system has no way of optimising its behaviour to achieve those goals.

- **Are there well defined constraints on the behaviour of the system? (critical)**

  AI systems have no common sense or moral compass. Without clearly defined constraints, they have no reason to adhere to any set of human values as they take actions to achieve their objectives.

- **Can we detect if the system is causing unintended side effects? (critical)**

  It is often the case that some impacts of the system's actions will be overlooked during development. It is critical that a feedback loop is established through on-going monitoring and open lines of communication with users to ensure that unforeseen harmful consequences can be quickly addressed.

- **Do we have a lot of (good quality) data? (recommended)**

  A key differentiator between AI systems (using machine learning) and traditional rule-based systems is AI's ability to identify complex or subtle patterns in data with little to no guidance from a domain expert. One price to be paid for this benefit is a reliance on large amounts of data.

- **Are we attempting to infer causal relationships? (caution)**

  AI systems that are used to make predictions in circumstances that are different from those under which its training data set was generated run the risk of under performing and leading to unintended consequences. Questions such as "What is the effect of

action X on person Y?" are causal in nature and may not be simply answered with AI systems that simply learn correlations (see section 3.2.3).

- **Does the environment change rapidly or suddenly over time? (caution)**

  AI systems are trained on historical data and so reflect the behaviour of the environment under past conditions. Sudden changes or drift in these conditions may lead to a deterioration in the system's performance.

- **Do we require the system's decisions to be explainable? (caution)**

  AI systems use observed correlations in the data to make their predictions. This is in stark contrast to how humans typically explain their decision making process, which generally involves some degree of causal reasoning. As a result, caution should be exercised if explainability is a key requirement of the system's legitimacy.

B. Harms Questionnaire

Below we describe an example of a harm questionnaire: a list of preliminary questions that can be asked about a specific AI system to obtain a qualitative sense of its overall level of risk with respect to harms it may cause on the people affected by it.

- **What is the purpose behind this system? Why was it built?**

  How can that purpose lead to harm for individuals, groups and society? For disadvantaged groups? Across gender, race, socio-economic status and other indicators of disadvantage?

- **Which behaviours will this system reward or encourage?**

  Behaviours of individuals, groups and society? Disadvantaged groups? Across gender, race, socio-economic status and other indicators of disadvantage?

- **How many people can be directly impacted by this system?**

  How is the breakdown across disadvantaged groups? Across gender, race, socio-economic status and other indicators of disadvantage?

- **On average across impacted individuals, how frequently does the system impact them? Hourly, daily, …, yearly?**

  How is the breakdown across disadvantaged groups? Across gender, race, socio-economic status and other indicators of disadvantage?

- **What's the absolutely worst type of harm an individual can conceivably be subject to as a result of a single interaction with the system**?

  This should be considered even if the harm is extremely unlikely to occur or has never occurred before. What are the groups most vulnerable to this harm? How disadvantaged are those groups?

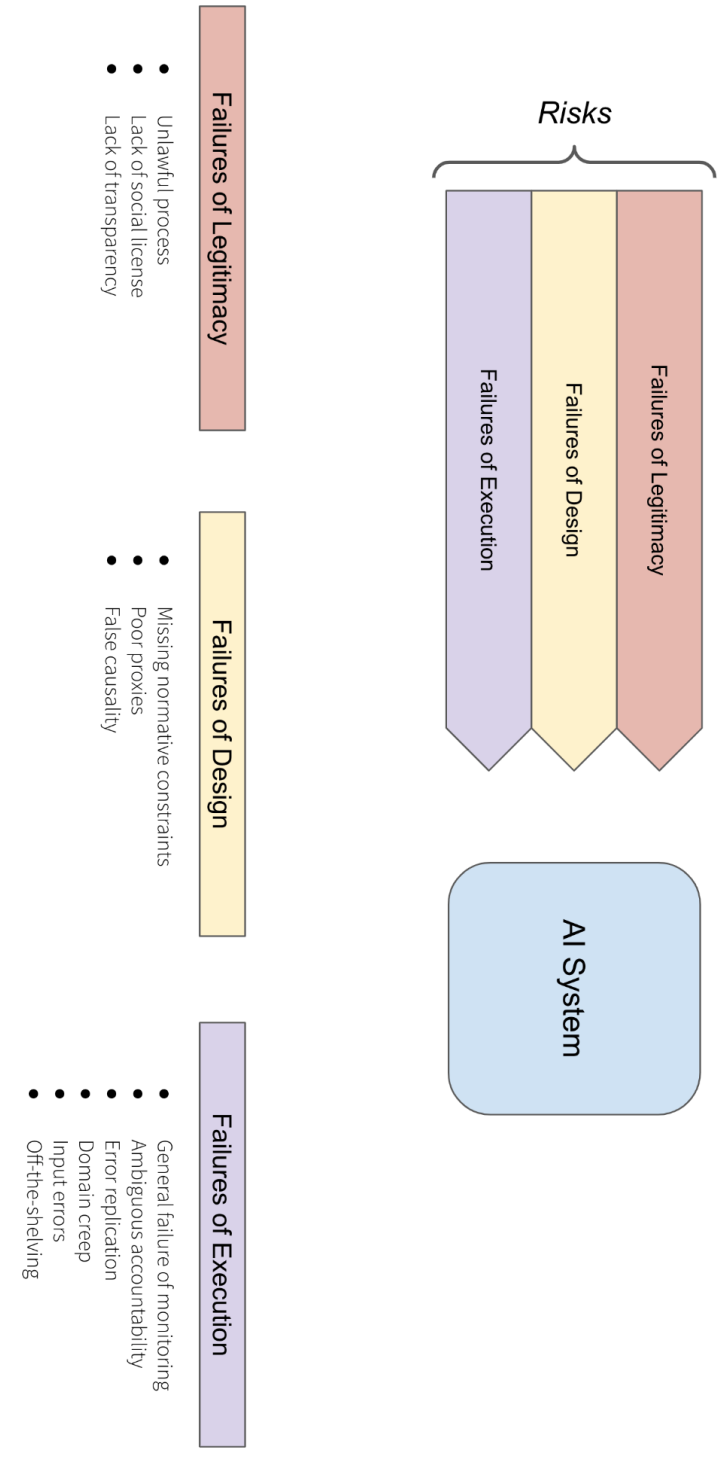
- **Same as above but for repeated, continuous interactions (instead of a single interaction)**.

- **What are the most significant harms the system can cause to an individual, as a result of a single interaction with the system**?

  This should be considered regardless of how likely the harms are to occur. What are the groups most vulnerable to each of these harms? How disadvantaged are those groups?

- **Same as above but for repeated, continuous interactions (instead of single interactions).**

C. Identified Risks



D. Suggested Actions