National Artificial Intelligence Centre

GRADIENT INSTITUTE

Implementing Australia's AI Ethics Principles:

A selection of Responsible AI practices and resources

June 2023

CSIRO Australia's National Science Agency

Citation

Alistair Reid, Simon O'Callaghan, and Yaya Lu. 2023. Implementing Australia's AI Ethics Principles: A selection of Responsible AI practices and resources. Gradient Institute and CSIRO.

Copyright and licence

This report has been prepared by Gradient Institute for CSIRO's National Artificial Intelligence Centre.

Authors

Alistair Reid, Simon O'Callaghan, Yaya Lu
(all at Gradient Institute)

Disclaimer

Gradient Institute and CSIRO advise that the information contained in this publication comprises general statements. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice.

To the extent permitted by law, Gradient Institute and CSIRO (including their employees and consultants) exclude all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

No warranty of completeness, accuracy or reliability is given in relation to the statements and representations made by, and the information and documentation provided by stakeholders consulted as part of the process.

Acknowledgements

The National AI Centre is funded by the Australian Government and coordinated by CSIRO.

This report has been prepared by Gradient Institute for CSIRO's National Artificial Intelligence Centre.

CSIRO acknowledges CEDA and Google as Foundation partners of the National AI Centre.

The authors would like to acknowledge:

- for feedback, discussions and contributions to specific sections of this report: Tiberio Caetano
- for regular discussions that helped shape the ideas reflected in this report, and for feedback on earlier versions of the report: Liming Zhu, Qinghua Lu, Tamara Ogilvie
- for supporting research and discussions: Dhivya Kanagasingam
- for feedback on drafts of this report: Aurelie Jacquet, Grace Abuhamad, Chris Dolman, Rita Arrigo, Stuart Banyard

The authors would also like to thank Fifth Quadrant for engaging in various discussions throughout the development of this report. In particular, we acknowledge their valuable contribution in conceptualising the visualisation presented in the Responsible AI practice infographic summaries.

We would like to express our sincere appreciation to other staff from Gradient Institute and CSIRO's Data61 for work that has influenced this report.

Photographs:

- unsplash.com: cover and pages ii, 1, 10, 13, 20, 44
- shutterstock.com: pages 4, 7, 9, 14, 19, 21, 22, 24, 27, 28, 33, 34, 37, 38, 43

Contact

National AI Centre
naic@csiro.au
csiro.au/naic
1300 363 400
+61 3 9545 2176

Gradient Institute
info@gradientinstitute.org
gradientinstitute.org

Contents

Executive summary

The National Artificial Intelligence Centre (NAIC) wants to support businesses to mature their Responsible AI (RAI) practices. To achieve this, NAIC has worked with Gradient Institute to provide an orientation to a selection of tools and guidelines that can support bridging the gap between the Australian AI Ethics Principles and the business practice of RAI.

The purpose of this report is to help raise awareness amongst AI practitioners of some important tools and guidelines that can be used to connect principles and practice. This is done in a two-step approach.

First, for each of the eight Australian AI Ethics Principles, we identify a pragmatic, but non-exhaustive selection of practices that promote the realisation of that principle. These range from broad activities such as "education on the responsible use of AI" to more specific ones such as "defining how to measure and monitor fairness using data". In total we propose 26 different practices relevant to Australia's AI Ethics Principles.

Second, for each of the identified practices, we

- explain what the practice involves, and how it implements an ethical principle
- identify who should apply the practice and when they should apply it within the AI system lifecycle
- orient the reader to the types of tools and guidelines available, so they can choose an approach that works best for their organisation
- point to impactful and popular examples of resources (including software tools and guidelines) and provide some guidance on how to differentiate them
- identify any gaps in resources to address the practice, and suggest courses of action to navigate them.

The space of applications for AI in businesses is rapidly evolving. AI can currently be used for a broad range of applications ranging from making personalised decisions at scale to having natural language interactions with customers. As the landscape of resources keeps evolving, current tools and guidelines will be supplanted by newer, more advanced ones. Some practices are likely to stay relevant and new practices will emerge. This implies that organisations should invest in enhancing their culture and governance processes in order to bring practices to a level of standard routine, while staying informed about emerging tools and guidelines and how they can aid in the execution of the various practices.

1 Introduction

Artificial Intelligence (AI) is increasingly used by industry and government organisations to make consequential decisions that affect people's lives. AI algorithms can be flexible and efficient at achieving business objectives, with capabilities such as making personalised decisions across entire cohorts, or having natural language interactions with customers. However, without being explicitly designed with the appropriate checks and balances they can also have unintentional negative impacts.

Businesses should care about using AI responsibly not only because it is the right thing to do, but also because it can help them prevent legal risk, avoid reputational risk, and establish trust with customers and stakeholders. A set of guiding ethical AI principles can provide a foundation for businesses to address the complex issues surrounding the use of AI. We structure this report around Australia's AI Ethics Framework and its eight voluntary principles of human, societal and environmental wellbeing, human-centred values, fairness, privacy protection and security, reliability and safety, transparency and explainability, contestability and accountability.[1]

It is important to recognise that these eight ethical principles exist as a tool to structure the users' thinking. Various frameworks offer ethical principles addressing the same set of concepts, organised around different themes such as justice and autonomy.[2]

For each of Australia's AI Ethics Principles, we identify some of the key relevant practices and orient the reader to resources to implement them. For each practice, we:

- explain what the practice involves, and how it promotes an ethical principle
- identify who should apply the practice and when they should apply it within the AI system lifecycle
- orient the reader to the *types* of resources available, so they can choose an approach that works best for their organisation
- point to impactful and popular examples (including software tools and guidelines) and provide some guidance on how to differentiate them
- identify any gaps in resources to address the practice, and suggest courses of action to navigate them.

It is anticipated that many of the businesses reading this report will be procuring AI systems from third-party vendors rather than developing in-house solutions. We caution that this does not absolve them of responsibility for how the system operates. The purchaser needs to be informed about (or ideally be involved in specifying) the system's algorithms, data and objectives. A client who simply trusts that a vendor has taken appropriate care — without applying their own due diligence — may be exposing themselves to uncontrolled risks for which they are ultimately accountable.

1.1 The need for Responsible AI in business

It is by now well established that the use of AI increases risk for businesses. The use of AI can create new sources of risk, as well as increase existing sources of both legal and reputational risk (for a detailed treatment, see Gradient Institute's report *De-risking Automated Decisions: Practical Guidance for AI Governance*[3]):

- **Legal risk.** The use of AI increases the risk that businesses fail to comply with laws and regulations, often without them realising. For example, AI can make discriminatory decisions that fail to comply with anti-discrimination law, or make inscrutable decisions in situations when the law requires those decisions to be explainable.
- **Reputational risk.** Even if the use of AI is legal in a certain context, it may cause reputational damage if it is considered controversial, unethical or untrustworthy by customers, regulators or the broader public. For instance, some technology companies have over the past few years discontinued many of their face recognition services in response to concerns raised by academics and the public society about the ethics of providing such services (in particular in light of the disparate performance of these systems across different demographic groups).[3]

Just like with human decision makers, AI systems must follow the relevant legal requirements, such as those related to anti-discrimination, privacy, and consumer protection. But these laws alone only provide a minimum standard for AI systems to follow – they are not sufficient to ensure that the AI system operates in alignment with the organisation's values, ethics and societal expectations.

1.2 Organisational approach to RAI

Organisations adopting **Responsible AI (RAI)** must contextualise ethical principles such as wellbeing, fairness or transparency to each AI system they create, and carefully balance these ethical goals against the system's business purpose. This will involve making practical commitments towards designing, deploying, maintaining and using AI systems in a way that is accountable to the people the AI system interacts with, minimises the risk of negative consequences and maximises the benefits to individuals, groups and the wider society. Thus, practices of RAI are designed to ensure that AI systems:

- operate in alignment with the organisation's ethics, objectives and constraints
- have a socially acceptable purpose, aligned with the views of people affected by the system, and of broader society
- do not cause unintentional or unjustified harm to individuals, society or the environment in the process of achieving their outcomes.

This report is structured around the Australian Government's Ethical AI Principles. It identifies varied resources that encourage alignment with each of these principles individually. However, it is important to note that numerous organisations have developed and published frameworks that offer salient advice and tools for bridging the gap between their own set of principles and AI in practice. [4] [5] [6] [7] [8] [9]

To deeply explore this body of literature, the AI Ethics Lab provides an interactive exploration of publications by industry and government organisations around the world today (reviewing over 100 distinct documents).[10] Senior directors navigating this space may consider:

- whether a document's ethical principles readily map to and provide coverage of their organisation's principles and values (and Australia's AI Ethics Principles)
- whether the guidance is practicable for their domain and organisation size – it may address an organisation using AI systems for a business purpose, but it may instead take a regulatory perspective, or assume the existence of ethics boards, research and development teams, or existing model policies that a smaller organisation may not possess
- what organisational changes are required to implement the guidance – it may take the form of lightweight tools (such as checklists and templates), or extensive overhauls of existing governance.

Furthermore, it should be noted that many organisations (particularly larger organisations) have established their own Responsible AI ethics committees and Responsible AI frameworks tailored to their values and use-cases.

1.3 Models, systems and accountable parties

A model's predictions have no impact until they are used to take actions, so understanding the degree to which predictive disparities or errors will map to harms and benefits requires contextual knowledge of the AI system and its use-case. In this report, we make the distinction between an AI system and an AI model as follows:

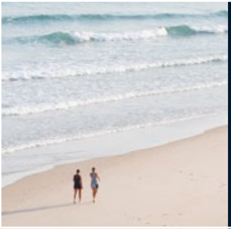
- **AI model:** A representation or structure that embodies the knowledge, patterns, or relationships gained by learning algorithms or other approaches, enabling it to make predictions, recommendations or to generate content.
  - *example: A model predicts how effective an offer will be at convincing a customer to purchase a product*
- **AI system:** A set of algorithms, models, interfaces, a pool of potential decisions or actions and a process for determining how to draw from this pool to achieve a specified objective (which may involve human decision makers and/or automated rules)
  - *example: A system chooses personalised offers to send to customers based on an effectiveness model, and delivers them via SMS*

Furthermore, within the text we will refer to three (broad) levels of roles within an organisation (see Section 9 for further details):

- The **system owner** role refers to the person (or persons) responsible for defining business and other objectives of the system in line with the organisation's strategy (as set by the board of directors), as well as for ensuring that the system implementation is fit for purpose and delivers on those objectives.
  - This is not fundamentally a technical role – where these decisions have ethical and technical components, it is expected that technical personnel will support the system owner in understanding the implications of, for example, design trade-offs or measurement choices.
- The **development team** is responsible for designing and implementing the AI system to meet the specified objectives and requirements of its owner.
  - This may include roles such as data scientists, software engineers or user experience experts.
  - In practice, some organisations may not do all the development work themselves, but instead draw upon a complex supply chain involving solution vendors (model, data, full solution providers) or machine learning service platform engineers.
- **Senior directors** are concerned with setting strategic goals and coordinating individuals within the organisation.
  - Leaders do not have direct oversight or make decisions at the level of a specific system, but are responsible for promoting Responsible AI in an organisation in terms of elements such as governance, policy and incentives.

RESPONSIBLE AI PRINCIPLE

Human, societal and environmental wellbeing



Throughout their lifecycle, AI systems should benefit individuals, society and the environment.

Elicit potential impacts

Account for the needs of all stakeholders while promoting inclusivity and equity.

DIRECTLY APPLICABLE TO:
system owners, development team

Assess impacts

Understand the positive and negative impacts the system's actions will have on people so they can be prioritised and managed.

DIRECTLY APPLICABLE TO:
system owners, development team

Set ethical objectives

Ensure that AI systems are explicitly designed to operate ethically (as well as meeting their business objectives).

DIRECTLY APPLICABLE TO:
system owners, development team

2 Human, societal and environmental wellbeing

Throughout their lifecycle, AI systems should benefit individuals, society and the environment.

AI systems, whether they are built to promote social good or serve a business purpose, must not create outcomes that unduly harm individuals, society or the environment. AI systems should instead have legitimate, defendable objectives, and any negative impacts should be accounted for throughout the AI system lifecycle.

This section points to key relevant practices for accounting for system impacts:

- **eliciting** the potential ways an AI system may impact people
- **assessing** an AI system against these impacts
- **balancing** mutually incompatible objectives in a responsible manner.

In addition to mitigating negative impacts, it is important to ensure that an AI system is having socially acceptable outcomes in the eyes of customers, affected stakeholders, and the public more broadly, which is examined in Section 3.

2.1 Elicit potential impacts

DIRECTLY APPLICABLE TO: system owners, development team

Most decision systems that allocate opportunities or resources or otherwise intervene in people's lives, including AI systems, will inevitably have positive effects on some people, and negative effects on others. It is unrealistic to expect that a "perfect" system can be created with beneficial outcomes for everybody.

Eliciting the perspectives of different stakeholders makes system owners aware of how the system's actions might affect them when specifying the system objectives and requirements. A common approach is to examine the system's intended purpose and proposed design, while brainstorming the range of impacts the system might have on various affected groups or individuals. During this stage, the goal is to identify as many potential impacts as possible rather than to decide how to prioritise them – this can come later after a more detailed analysis of their likelihood and severity has been conducted.

This practice is primarily led by the system owner, who needs to contextualise the system's impacts to subsequently assess them. The system owner cannot undertake this task alone, however, as the effectiveness of any identification exercise relies on the diversity of perspectives and expertise involved. Consulting affected stakeholder representatives, legal experts and domain experts can help anticipate a broader set of problems and is therefore beneficial to begin as early as possible in the AI system lifecycle (often before the system is designed, or as part of its design). Here it is important to clearly inform participants of the use-case that the system is being considered for, and the precise nature of its potential actions, to identify relevant concerns and document the scope of analysis.

Resources for eliciting impacts

- Diverse Voices[11] templates a workshop to elicit diverse perspectives about the impacts of a technology policy, which would generalise effectively to system impacts.
- The Ethics Canvas[12] is a lightweight resource that is well suited to facilitate a workshop about the harms and benefits of the system. It provides freely–accessible templates for the user to conduct workshops to explore the harms and benefits of a system, prompting participants to think about the key areas of impact an AI system may have, and who may be affected.
- Microsoft's Judgement Call[13] is a freely accessible kit for the user to conduct an internal team-based activity in which participants roleplay the system's affected stakeholders who are tasked with providing product reviews. The intent is to cultivate empathy in the development team for the users and uncover harms through the shift in perspective.
- An ethical matrix is a tool for system owners and developers to brainstorm a list of affected stakeholders and a set of potential issues that concern them. It involves constructing a table mapping stakeholders (on one axis) to issues that concern them (on the other axis). This serves as a starting point for identifying who should be brought into the conversation to express their concerns about potential impacts.[14]

Checklists offer a complementary approach to workshops as a means of identifying impacts.

- The Ethical OS Toolkit[15] provides a checklist to explore eight areas of risk and social harm often attributed to new technologies. Illustrative scenarios provide advice on what to do when the system owner and development team encounter them.
- The Ethics Guidelines for Trustworthy AI[16] includes a pilot checklist which categorises high-level questions under the European Commission's seven key requirements for trustworthy AI systems (which align closely to Australia's AI Ethics Principles in structure and coverage).
- The Azure Application Architecture Guide's "Types of harm" section prompts the user to think about many different concerns such as physical injury, emotional harm, economic loss, privacy loss and manipulation [17]

2.2 Assess impacts

DIRECTLY APPLICABLE TO: system owners, development team

Assessing a system's impacts on people, society and the environment helps the owners specifying the system and the developers building the system to understand the positive and negative effects the system's actions will have so that they can be prioritised and managed.

Algorithmic impact assessments provide a process to document the magnitude and likelihood of an AI system's harms and benefits, which allows system owners to gauge whether the system is of net benefit to its owners and to society, and decide how to specify system objectives and requirements to control them. These assessments are best initiated prior to development, as they can help with assigning accountability for aspects of the system's performance, as well as specifying mitigation strategies that will be implemented prior to deployment.

However, it is also important to conduct ongoing or periodic assessment to catch unforeseen impacts once the system is deployed, as the way the product or service is used, or the patterns it learned from historical data may change over time.

Many organisations have published guidelines or templates for impact assessments. The following are some examples:

- The Canadian Government has an algorithmic impact assessment that is mandatory for procurement of their internal systems.[18] They use an online scorecard tool aimed at identifying harmful impacts that arise over the lifecycle of AI systems, without requiring the user to have deep RAI expertise.
- Microsoft's Responsible AI Standard[19] includes an impact assessment template and an extensive user guide that details activities and guidance to help complete the assessment.
- AI Now Institute's Algorithmic Impact Assessment[20] provides step-by-step guidance for conducting an algorithm assessment (predominantly aimed at public agencies but adaptable for use in the private sector).
- ISO 42005 standard, specific to impact assessments, is currently under development.[21]

Considerations:

- Impact assessments need to consider many affected groups or individuals, including the broader public, regulators, people who bear the system's impacts, people who operate the system, are responsible for the system, or even whose jobs are displaced by the system.
- Impact assessments can be conducted and updated at any point during the system's lifecycle although it is advantageous to begin early to reduce costs associated with mitigation, remediation or reputational damage.
- Impact assessments are only valid for the specific use case they examine. It may be that an organisation wishes to use the same AI technology in multiple applications (such as computer vision for inventory management and store surveillance). Each potential use-case needs to be assessed, and the assessment's scope documented).
- The potential for incorrect use must also be considered. For example, if a user queries a large-language-model chatbot as an alternative to using a search engine, the results may seem credible, without necessarily being appropriate, truthful or complete.
- Vendors of AI products and services may apply some controls around how their technology is used[22], but many do not. As discussed in the introduction, a client who assumes that a vendor has taken appropriate care — without applying their own due diligence — may be exposing themselves to uncontrolled risks for which they are ultimately accountable.
- Standard measures (such as accuracy) focus on system errors and assume they are all equally important. In reality, some errors may be more impactful on average than others, and the same errors may be more impactful for specific individuals or groups than others. It is important to choose metrics that align with identified harms, and if a specific harm can't be meaningfully quantified, to identify what steps might be appropriate to mitigate it.
- When a system has been assessed, the owner can decide whether it is fit for purpose. However, it is a common pitfall to critique an AI system against a "perfect" system that has no negative impacts, when the key question is whether the benefits outweigh the harms in the eyes of the owner, and the public more broadly. A more realistic point of comparison is a feasible alternative such as a pre-existing system, manual decisions, simple rules or even not deploying the system. Various guidelines discuss baseline selection. [23] [24] [25]

2.3 Set ethical objectives

DIRECTLY APPLICABLE TO: system owners, development team

Having identified potential harms, the owners of AI systems may specify multiple design objectives, including business objectives such as profitability or accuracy, alongside ethical objectives such as fairness or transparency. Making ethical objectives primary objectives ensures that an AI system is designed to operate ethically as well as meet its business objectives. However, design trade-offs in AI systems with multiple objectives are inevitable because:

- design steps to prioritise a given metric may come at a cost to others (e.g. due to algorithmic and data limitations)
- system objectives may be inherently competing or conflicting (such that it is not feasible to align them).

Deciding how to balance competing objectives requires ethical — rather than technical — judgement, especially when a system's impacts cannot be easily compared or measured against one another. For example, is it better to have a system that uses no personal data, or a system that uses personal data but makes fewer errors?

When building an AI system, the key design trade-offs need to be identified by the development team and communicated to the system's owner. The owner then needs to decide how to prioritise and balance different aspects of the system's performance to align it with the organisation's values, objectives and constraints.

Resources to identify design trade-offs

AI models are usually designed to maximise accuracy, assuming this will align with most system objectives. But it is also possible to directly optimise measures of impact, as is done in cost-sensitive learning.[26]

The drivetrain approach offers a general and flexible way to adjust a system's parameters to optimise a specified objective.[27] Likewise, many fairness toolkits provide the functionality to improve a specific fairness objective (see Section 4).

These methods typically involve identifying (or designing) levers or parameters that control the performance of the system. While they are intended to improve a specific objective, they may affect other system metrics positively or negatively. Here, identifying and characterising the trade-offs between competing objectives is a technical problem that could be naively approached by manually exploring the effect of model parameters, but when models are expensive to retrain or have many parameters, exploration may require more sophisticated methodologies.[28]

Approaches and challenges for trade-offs

To prioritise and balance different aspects of the system's performance to align it with the organisation's values, objectives and constraints, a system owner needs to be able to steer the system's design decisions.

A simple approach is to use *satisfactory metrics.*[29] This involves identifying a "primary" objective and converting the others into constraints. For example, a development team may optimise profitability or accuracy while maintaining an acceptable level of a specific notion of fairness. This approach provides ethical safeguards but does not optimise the ethical performance to the same extent as the business performance.

An alternative strategy is to assign 'price' or exchange rate to different system objectives such that they can be aggregated - a process called 'scalarization'.[30] This allows developers to target a single measure that encodes a prioritised combination of the objectives. However, specifying prices is often as difficult, if not more difficult, than deciding how to balance the objectives. It is often more intuitive for a system owner to examine available options and choose between them than it is to specify how much a unit of fairness measure is worth. In doing so, they may also over-simplify nuanced requirements and preferences.

Decision support tools can help system owners to interpret and balance complex competing objectives. Preference elicitation tools from other fields such as economics or operational research can be applied to this task, as demonstrated by the AI Impact Control Panel, a proof-of-concept decision-making tool developed by Gradient Institute with support from Minderoo Foundation.[31]

Open challenges exist in eliciting preferences from human decision makers:

- The complexity of trade-off decisions increases with the number of criteria. Decision makers may prefer tools that ask them to consider only two criteria at a time.
- A human decision maker's preferences may depend on the order in which options are presented to them. It is difficult to determine if a tool has guided a user to their "ideal" trade-off except under simplifying assumptions.
- A human decision maker's judgement may suffer from cognitive biases such as loss aversion (placing more importance on sacrificing performance than gaining performance). The NAUTILUS method from the literature aims to address this by starting from a relatively weak position and exploring a series of incremental improvements.[32]

RESPONSIBLE AI PRINCIPLE

Human-centred values

Throughout their lifecycle, AI systems should respect human rights, diversity, and the autonomy of individuals.

Design for human autonomy

Respect and preserve human agency and decision-making capabilities.

DIRECTLY APPLICABLE TO:
development team

Achieve outcomes ethically

Justify the means by which outcomes are achieved.

DIRECTLY APPLICABLE TO:
system owners

Incorporate diversity

Promote systems that benefit the broader community and mitigate against perpetuating the implicit biases of its owners and developers.

DIRECTLY APPLICABLE TO:
senior directors, system owners, development team

3 Human-centred values

Throughout their lifecycle, AI systems should respect human rights, diversity, and the autonomy of individuals.

AI systems should be designed with the diversity, autonomy and fundamental rights of individuals in mind. In this section, we suggest relevant practices for ensuring an AI system respects these human-centred values:

- designing for **human autonomy**
- justifying the **means** by which positive impact is achieved
- incorporating **human diversity** into the design of the system.

3.1 Design for human autonomy

DIRECTLY APPLICABLE TO: development team

Autonomy refers to the ability of a person to self-govern. Threats to this include practices such as deception, unfair manipulation, and unjustified surveillance.[1] The avenues by which AI systems can manipulate people are greatly expanded by the arrival of generative models that can generate convincing content that appears to be created by a person. However, even well-intentioned systems can impact an individual's autonomy by removing or limiting their freedom of choice.

The first step to promote autonomy is to provide appropriate transparency and explainability of AI systems to end users or consumers (see Section 7). Businesses can further promote autonomy by exposing control mechanisms to users, avoiding misleading "dark patterns" in interface design[33 34 35] and asking for consent before processing a person's data or making decisions for them.

Individual control over an algorithm

A first step towards improving autonomy is to give users some level of control over the algorithm. A simple example is the thumbs-up button found in many recommender systems such as online music and video streaming platforms.

More sophisticated controls can be found in, for example, Google's advertising settings.[36] It should be noted that offering more complex controls may require users to possess relevant technical understanding to make use of them.[37] Google's *Feedback + Control* guide[5] provides guidance on how to design control and customization interfaces for service users.

Motivation is also an issue. If controls are too difficult to use, or the user has too many options, they may become fatigued and disengaged. This could lead to a user accepting the system's default settings rather than exercising their autonomy.[38 39 40]

Establishing user consent

Establishing consent for how a person's data is collected or processed is essential to respecting their privacy and autonomy and may be subject to legal requirements. It is important to obtain legal advice on how to do so for a given application. In many settings, a person who does not consent to specific functionalities of the system may be able to continue engaging in a more limited capacity. For example, a person receiving advertising might request to have generic / non-personalised advertising shown to them. Furthermore, to make an informed decision about whether they want to grant consent, impacted persons need to be aware of who is handling their data, what they are recording, and for what purpose (see Section 7: Transparency and explainability). Various web consent-management services offer best-practice guidance specifically for web applications.[41 42 43]

The right to opt-out of automated data processing has been recognised in some settings by, from an Australian perspective, foreign regulations including the General Data Protection Regulation (GDPR). Australian businesses of any size may be required to comply with the GDPR if they conduct aspects of their operation within the European Union – seek legal advice where necessary.[44]

3.2 Achieve outcomes ethically

DIRECTLY APPLICABLE TO: system owners

When establishing the legitimacy and social licence of AI systems, it is important to consider not only the positive impacts but also the means by which they are achieved.

Justifying AI systems with respect to the wellbeing principle (Section 2) requires arguments that they create positive impact. However, systems that create positive outcomes through unscrupulous methods are not necessarily ethical or socially acceptable.[3] A human-centred values approach to AI system development should consider fundamental international and Australian human rights conventions, such as the right to equality and non-discrimination.[45]

Such concepts are often less tangible than other qualities of AI systems like accuracy or reliability and may require human judgement. Short questionnaires like the "Should We?" test found in Westpac's code of conduct[46] can be useful to determine whether a system might meet societal expectations, where questions like "Would you be comfortable explaining the ideas behind the system to a friend or family member?" can help identify aspects of the system that may fall short of social standards or potentially impact human rights. The Organisation for Economic Co-operation and Development (OECD) has developed a framework for the classification of AI systems which supports risk assessment and risk management.[47] However, it is necessary to recognise that these assessments are developed to meet existing regulations or governance requirements, and are not necessarily designed with Responsible AI in scope.

We note that in terms of respecting laws and regulations, Australian laws and regulations should be the first and foremost concern for Australian organisations, who may then raise the bar by choosing to also adopt practices from other parts of the world. However, there are implications in doing so relating to challenges of compliance with multiple regimes, and associated compliance risks.

The growing use of generative AI raises new questions around data use, such as how the use of copyright materials in the large corpuses of training data required by these algorithms might infringe on the creator's intellectual property, particularly if the model were to inadvertently plagiarise these works.[48]

A broader approach to assessing the social licence of a system is to engage in public consultation. These consultations could take the form of market research or focus groups and can also act as a forum to identify potential harms, as discussed in Section 2. The Australian Government Department of the Prime Minister and Cabinet (PM&C) provides a *Best Practice Consultation* publication[49] with guidelines for policymakers, including recommendations that can be generalised more broadly.

Social contracts for the use of AI systems (for example, AI for Social Good's *The Social Contract for AI*[50] and AI World Society's *Social Contract for the AI Age*[51]) are resources for companies to publicly demonstrate commitment about how they will (and won't) use AI. Further information on the importance of disclosure and involving the public in drafting impact assessments can be found in Section 7.

3.3 Incorporate diversity

DIRECTLY APPLICABLE TO: senior directors, system owners, development team

Humans are diverse, therefore there should be diversity in every part of the AI system lifecycle.[52] This includes the choice of which stakeholder and user groups to consult, the choice of the data that is fed into the system, and the composition of the development teams themselves.

AI systems designed without diverse and inclusive practices can be perceived as untrustworthy, unfair or actively discriminatory, leading to the services being ostracised by the community, or to people with diverse abilities disengaging.[53] [54]

Conversely, including a diverse set of perspectives when designing a system benefits the broader community and mitigates against perpetuating the implicit biases of a system's owners and developers. AI services actively adopting diversity practices are more likely to receive public appraisal. For example, although AI facial recognition technology is generally seen as potentially risky and harmful in most settings, for the specific use-case of assisting the blind and visually impaired community with day-to-day activities it has been positively received.[55] [56] [57]

Resources for diversity and inclusion in AI

The *Diverse Voices* guide by the University of Washington's Tech Policy Lab templates a method to elicit perspectives from under-represented groups about the impacts of a technology policy. This method would generalise effectively to discovering and evaluating system impacts on a diverse cohort.[11] To make these diverse voices count it is necessary to ensure the system designers use these perspectives to better include people with diverse backgrounds and abilities to benefit the broader community.

Partnership on AI is a non-profit group that brings together diverse voices from across the AI community to develop actionable resources for AI. They develop best-practice documents, undertake studies on challenges to diversity in AI, and host workshops on inclusive research and design for AI practitioners.[58]

OCAD University in Canada has published The Inclusive Design Guide, which is designed for a general audience and can be applied to workshops, meetings, conferences, services, and physical products. The guide includes practices, tools and activities to promote inclusive design.[59]

Microsoft's In Pursuit of Inclusive AI[60] document offers best practice guidance for inclusive AI design, covering topics such as:

- avoiding dataset biases that can perpetuate cultural bias
- enlisting customers to correct fairness
- re-engaging disadvantaged demographics through privacy and consent
- hiring staff from diverse backgrounds, disciplines and demographics.

CSIRO's National AI Centre Think Tank article[52], Deloitte's *Opening doors of opportunity* report[61] and opinion pieces like *Human-centred AI to build a trustful customer experience in retail*[62] are some examples of the many resources available that highlight the importance of diverse engagement, both internally and externally to the system.

RESPONSIBLE AI PRINCIPLE

Fairness

Throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.

Contextualise fairness

Consider what is fair in the context of the system's impacts and who is affected.

DIRECTLY APPLICABLE TO:
senior directors, system owners

Measure fairness

Quantify fairness concerns to promote effective oversight and mitigation.

DIRECTLY APPLICABLE TO:
system owners, development team

Mitigate unfair impacts

Reduce the risk of the system introducing, perpetuating or amplifying societal inequalities.

DIRECTLY APPLICABLE TO:
system owners, development team

4 Fairness

Throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.

When a system's decisions have human, societal or environmental impacts, some people may miss out on their share of the benefits, or disproportionately bear the harms. Unfair impacts can arise from historical disadvantage or discrimination reflected in the data[63] [64], systematic bias in how the data was curated or recorded[65], or how the AI components of the system were specified.[66] [67] Educational materials such as interactives explaining the concept of algorithmic fairness may be helpful for readers who are not familiar with the common measures of fairness and their interpretations.[68]

Fairness problems tend to be accidental, and a system's operators are usually unaware of them. In this section, we provide a series of resources to effectively address fairness by:

1. **contextualising** fairness to the system's operational use-case
2. defining how to **measure and monitor** fairness using data
3. adopting **mitigation strategies** to improve the fairness metrics in conjunction with other relevant system metrics.

4.1 Contextualise fairness

DIRECTLY APPLICABLE TO: senior directors, system owners

While it is broadly accepted that fairness is important, what constitutes fair outcomes or fair treatment is open to interpretation and highly contextual. What constitutes a fair outcome can depend on the harms and benefits of the system as well as one's own internal value system. It is the role of the system owner to consult relevant domain experts, affected parties and system stakeholders to determine how to contextualise fairness to their AI use-case, informed by foundational guidance and organisational values set out by the senior directors of their organisation.

Resources for contextualising fairness

- Tools for eliciting impacts including Diverse Voices[11], The Ethics Canvas[12], Ethical OS Toolkit[15], Microsoft's Judgment Call[13] and the algorithmic impact assessments discussed in Section 2 also examine how individuals and groups might be disparately impacted.
- Microsoft's AI Fairness Checklist[69] provides key points of consideration for the system owner and development team through the system lifecycle (where most contextualisation occurs in the 'Envision' and 'Define' phases).
- FEAT Fairness Principles Assessment Methodology (Monetary Authority of Singapore's Veritas consortium)[7] poses questions to system owners to guide them through the process of contextualising fairness and specifying associated measures and controls.
- Using AI to make decisions: Addressing the problem of algorithmic bias (Australian Human Rights Commission)[45] explains how inequality makes its way into algorithmic decisions in terms of societal, data, and algorithmic contributions and identifies types of mitigations (see 4.3).

4.2 Measure fairness

DIRECTLY APPLICABLE TO: system owners, development team

Quantifying fairness is an important step in enabling an AI to minimise disparate impacts. A wide range of metrics have been developed by the Responsible AI community to address various concepts of fairness. Faced with a catalogue of off-the-shelf options, AI practitioners generally have two options (and may apply both):

- evaluate many metrics, to decide whether they reveal problems with the system; or
- specify metrics to measure identified harms.

The former provides a safety net when system owners are uncertain which to focus on. For example, automatic assessments such as the Aequitas Bias and Fairness Audit Toolkit[70] can be leveraged to provide a digestible summary of standard metrics and explain why they might indicate a problem with the system.

The latter provides a more focussed approach where the system can be specifically designed to measure and mitigate relevant harms through its lifecycle. Resources for this approach need to assist users in specifying fairness measures that adequately capture their intent with relation to the identified harms.

Resources for choosing fairness metrics

Metric selection trees (such as the Fairness Compass[71] or Fairness Tree from the Aequitas Bias and Fairness Audit Toolkit) guide the user through a sequence of targeted questions about their AI use case to arrive at a suggested metric. However, the approach must be exercised with caution:

- The wording of these tools is necessarily precise, but may be difficult for a non-technical user to contextually interpret.
- The procedural nature can make a decision seem clear-cut, but in practice different perspectives will often lead to different metrics. It may be beneficial to apply a tool multiple times focussing on different concerns.
- The process may not adequately capture the reasoning behind a decision. For example, if trusted ground truth is not available, the Fairness Compass directs users to demographic parity not because it necessarily encodes their intent, but because it is the only metric supported by the data.

A more flexible and open-ended approach, on the other hand, is for system owners to engage with the system developers to translate their intent into metrics. Although documentation around metrics uses precise, technical-sounding language, it is essential that system owners aim to fully understand their metrics, and are aware of what is lost in translation when a fairness concept is represented by a measurement. For example, the FairLearn User Guide[72] warns of:

- **solutionism:** assuming the best solution is a technological one
- **formalism:** focusing on measurable aspects of fairness that don't fully capture the nuance of the problem
- **framing:** where the scope of the impact considered is too narrow.

In terms of familiarisation with the academic debates around fairness, the FairLearn User Guide debates two commonly-adopted fairness metrics: demographic parity and equalised odds. Along similar lines, a What-If Tool blog post provides perspectives from 5 fictitious experts exploring the key differences of 5 different fairness metrics.[73] In terms of choosing metrics, we identify a number of remaining challenges:

- The rapid growth of this field has led to inconsistent motivations, terminology and notation, which makes cataloguing and comparing resources difficult, as addressed in a recent review.[74]
- Most guidance is concerned with classification tasks, while systems that utilise clustering, recommender engines or even regression models are under-served.
- Approaches to quantifying ethical issues such as language bias are still maturing.

Resources for evaluating fairness metrics

There exists a plethora of toolkits to evaluate commonly-used fairness metrics. When deciding which tool to employ, key considerations include whether a toolkit:

- evaluates the desired metrics
- offers mitigation techniques targeting the desired metrics
- is interoperable with existing workflows and data structures
- provides adequate guidance around selecting and interpreting metrics.

For systems developed in Python's de-facto data science software stack[75], the dominant offerings in this space are AI Fairness 360 (AIF360)[76] and FairLearn. These are software libraries with flexible and comprehensive functionality:

- AIF360 has the larger catalogue of state-of-the-art metrics and mitigations, but demands a high level of expertise to effectively use them.
- FairLearn's documentation provides clear standalone guidance about which methods are appropriate for various tasks and their limitations, while AIF360 users are often pointed to academic literature to understand a given method's purpose.
- Although their functionality overlaps significantly, Fairlearn includes regression fairness metrics, while AIF360 supports multi-class classification and individual fairness metrics.
- Both tools are open source, accept code contributions from the community, and support their users in the form of tutorials and community spaces (Slack for the AIF360 community[77], and Discord for Fairlearn[78]).
- AIF360 offers R support, in addition to Python.

Developers in environments other than Python or R may face an interoperability barrier requiring them to export data in order to analyse it. An exception to this pattern is machine learning platforms that bundle RAI functionality such as:

- Amazon SageMaker (SageMaker) Clarify[79], which offers functionality to evaluate and monitor group fairness and model reliability within the SageMaker cloud machine learning platform.
- Microsoft Responsible AI Scorecard[80], which generates model reports from Azure Machine Learning studio containing group-level error distributions, along with model feature importance plots.
- Salesforce *Einstein Discovery*[81], which enables users to declare certain data variables as protected attributes, check for correlations with other variables and observe differences in outcomes between groups.

However, these tools are less flexible and comprehensive than the AIF360 or Fairlearn libraries. Users looking for a more interactive (and somewhat more technical) experience may consider:

- Google's What-If Tool[82] can be used to explore the effect of changing data or model parameters. The What-If Tool can be used with Tensorboard or via the Tensorflow (Python) library.
- The Aequitas Bias and Fairness Audit tool provides automated assessment of a system against a range of standard measures. This tool is not tied to a specific workflow, but involves importing CSV format into a web interface (the service can be hosted locally to avoid sharing data).

A comprehensive review of these and many other resources can be found in *Landscape and Gaps in Open Source Fairness Toolkits*[83] and CSIRO's Responsible AI Pattern Catalogue[84]. While the community is doing a commendable job of supporting practitioners, it is important to be aware of the limitations of available resources:

- Most resources focus on group fairness, which compares average outcomes or performance across groups, such as across gender, races, or socio-economic statuses. Resources largely ignore fairness on an individual level, which is concerned with whether people with similar attributes receive the same outcomes. This is likely due to the philosophical and technical challenges posed when comparing individuals without abstracting to a few characteristics. AIF360 is the exception, offering rudimentary individual fairness inequality metrics.
- Standard fairness metrics examine actions or errors, and rarely account for long-term impacts. For example, lowering the credit-score cutoff for a disadvantaged group to attain a loan — to provide them with the same access to the product as the general population — may result in a high default rate that ultimately harms them in the long term.
- Evaluating group fairness metrics requires the use of individual-level protected attribute data. This data may not be available, and even if it is, collecting or retaining information against certain protected attributes may pose unacceptable legal, privacy or reputational risk. In lieu of recording protected attributes, inferring them also poses its own set of risks.
- Fairness metrics that examine errors require trusted ground truth of the "correct" decision. If this ground truth reflects extrinsic fairness problems, the metrics will be blind to them.

4.3 Mitigate unfair impacts

DIRECTLY APPLICABLE TO: system owners, development team

If a system's fairness performance is of concern, the owner may use various approaches to reduce the risk of the system introducing, perpetuating or amplifying societal inequalities.

Outright withdrawing the system from service may be an option, but may itself be harmful and generate unfair ethical impacts, especially if the system was performing a beneficial or essential role.

Intuitive safeguards — such as withholding protected attributes such as gender from the system — offer false security, and may even exacerbate an existing problem.[85]

A more flexible and effective approach is for the system owners to ask the development team to apply one of the mitigation strategies below during the system's design and development, and work with them to strike a deliberate balance between the system's ethical and business objectives that achieves acceptable performance for both.

Resources for improving system fairness

Fairness mitigation tools generally fall into three categories:

- **dataset pre-processing:** the dataset is transformed to reduce correlations between protected attributes and the attribute that the system is trying to estimate. If the same dataset is used in multiple applications, this may be an efficient way to improve fairness across all of them, provided they have compatible fairness requirements.
- **algorithmic in-processing:** the model is trained to optimise a fairness metric as well as other performance metrics such as accuracy. This type of mitigation specifically requires a training process where the learning objective can be modified. A popular approach for neural networks is adversarial learning with fairness objectives.[86]
- **decision post-processing:** the decisions (which are based on model predictions) are adjusted to improve the fairness metric. This approach is straightforward to implement, although it usually requires the system to collect protected attributes to use in its decision process. This approach closely aligns with positive discrimination, a practice that may be controversial in some settings.

Resources that provide these algorithmic mitigations include:

- FairLearn, which provides 6 strategies covering a range of pre-processing, post and in-processing approaches. The user guide documentation offers clear information about which techniques are compatible with classification or regression, and which fairness metrics they can target for improvement.
- AIF360, which includes 10 state-of-the-art mitigation strategies from academic publications. However, the toolkit offers little clarity about what they do or how to choose between them, so there may be a difficult learning curve in determining which tools are appropriate and effective.
- The scikit-lego *Fairness* module[87] (whilst small) contains scikit-learn compatible models with equal opportunity or demographic parity constraints.
- Google's What-If Tool allows the user to tune decision thresholds interactively as a form of post-processing.

An AI practitioner may face challenges in leveraging these resources:

- Teams using off-the-shelf models have limited mitigation options if the vendor has already committed to the algorithm's design or training data. If system owners are not able to bring their own data, measures and objectives to the engagement, they may still be able to apply post-processing mitigations, provided that the vendor's model provides confidence scores.
- Some tools require information about the users' protected attributes to function. This may violate existing user agreements relating to data use and data privacy.
- A task has to be framed in terms of classification or regression to apply these approaches.
- Approaches for moderating generative outputs are still evolving. The key challenge is developing methods for flexible and reliable automated moderation.
- Toolkits don't have mitigations to target every metric they can measure (although addressing supported measures may indirectly improve an unsupported metric).
- Prioritising a given fairness objective will necessarily de-prioritise other objectives. (Balancing competing objectives is a challenge discussed in Section 2).

Guidelines on algorithmic bias from the Australian Human Rights Commission[88] decompose algorithmic bias into extrinsic, data and algorithmic components, and use this decomposition to derive a taxonomy of potential mitigations. This points to other mitigation strategies:

- Developers can explore how to acquire or record more appropriate data for datasets that are outdated, contain insufficient relevant information or under-represent some individuals, which can lead to inaccurate outcomes in an AI system.
- Some fairness problems can be mitigated by rethinking how the target is defined, since machine learning tasks usually specify a predictive target to quantify abstract concepts like profitability, creditworthiness or job suitability. The degree to which the target is an accurate representation of the true concept may differ across groups and circumstances.

RESPONSIBLE AI PRINCIPLE

Privacy protection and security

Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.

Consult privacy and security experts

Ensure established legal and technical practices are applied.

DIRECTLY APPLICABLE TO: senior directors, system owners

Guard against attacks

Ensure that malicious actors cannot manipulate or compromise the system or its data.

DIRECTLY APPLICABLE TO: development team

Minimise the collection of personal information

Only use the most relevant data records and attributes to reduce risk of privacy exposure.

DIRECTLY APPLICABLE TO: system owners, development team

Consider using privacy preserving models

Use algorithms that minimise privacy exposure by design.

DIRECTLY APPLICABLE TO: development team

5 Privacy protection and security

Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.

Privacy protection and security are essential considerations for any business that uses data about its customers, employees or community, whether the dataset is ingested by an AI system or used for any other business process. Furthermore, a business that fails to meet its privacy and security obligations may incur significant reputational damage and face legal penalties.

It is important to recognise that privacy and security are two distinct yet interconnected fields of expertise (for instance, a security breach that leaks personal information is also a privacy breach). A widely accepted model of security (in the context of information security) is the confidentiality, integrity, availability triad:[89]

- **Confidentiality** is related to protection against unauthorised access of data.
- **Integrity** is related to protection against data being tampered with or manipulated.
- **Availability** is related to protection against data or systems being inaccessible.

Privacy, on the other hand, is related to appropriate use of data in terms of when it is used, for what purpose and with what permissions.[90] When a system utilises algorithms and personal data, it may be subject to specific legal requirements such as data protection acts.[91] [92] [93]

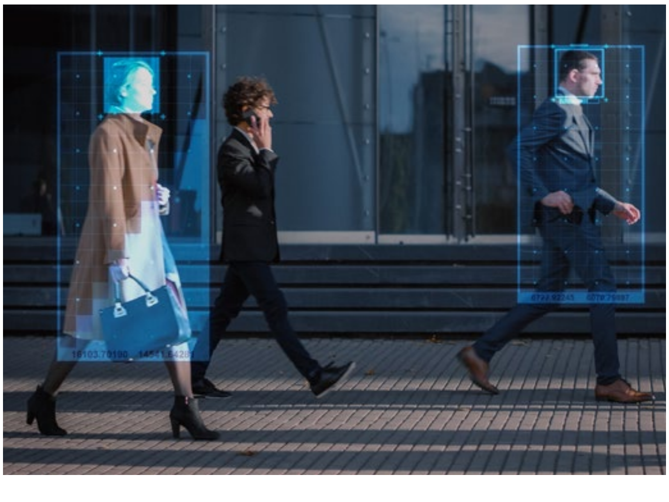
A key resource for organisations to address privacy risk and security risk is consultation with specialists in each field. This section will also discuss a limited set of practices that are particularly relevant to AI design and implementation, and might be included within a broader set of organisational security and privacy practices. These relate to guarding an AI system against attacks and designing its data processing in a way that can reduce harm in the event of compromised security. Assessment templates, such as those provided in AI Privacy 360[94], may also help the system owner to document and prioritise their system's risks and mitigations.

5.1 Consult privacy and security experts

DIRECTLY APPLICABLE TO: senior directors, system owners

Unlike the other Responsible AI principles, privacy and security are both fields of expertise with established legal and technical practices beyond the scope of this report. Security and privacy experts can be engaged to provide a range of resources such as reviews, assessments, advice and training.[95] [96]

Australian organisations should take note that Australian privacy laws are currently under reform. In their Privacy Act Review Report,[97] the Australian Government Attorney-General's Department put forward 116 proposals, many of which are relevant to organisations using AI. Examples include new data subject rights, new requirements around the use of personal information for automated decision-making and increased powers for regulators overseeing and enforcing the Privacy Act. These proposals — and the feedback received on them — are expected to shape future legislation. To find out more about this reform, commentary from legal experts could provide a valuable resource.[98] [99] [100] [101]



5.2 Guard against attacks

DIRECTLY APPLICABLE TO: development team

An AI system can be vulnerable to attacks from malicious actors, either through the design of the system, or through compromised security of its data or execution environment. Guarding against these attacks helps towards protecting the confidentiality, integrity and availability of the system and its data.

The level of exposure to attack depends on multiple human and IT security factors:

- If detailed knowledge about the algorithm and its parameters is disclosed or stolen, attacks can be devised to specifically target the intricacies of how the model works (**white-box attacks**).
- If the interface allows attackers to probe the model with strategically constructed inputs, this exposes the system to attacks that can infer the model's training data, estimate how the model works, and/or search for inputs to manipulate it from examining the inputs and outputs only (**black-box attacks**).
- If the integrity of the data the model is trained on is compromised, then an AI system can be manipulated to learn incorrect or compromised behaviours.

In order to defend against these types of attack, system developers can consider using adversarial robustness tools to investigate the resilience of their models. For example, the Adversarial Robustness Toolbox 360[102] (art360) provides guidance and categorises potential attacks into 4 categories:

- **evasion:** applying small changes to an input in order to manipulate the output without appearing obviously anomalous
- **poisoning:** corrupting the training data to degrade performance or manipulate model outputs to achieve a desired goal
- **inference**: querying a model to reconstruct its training data. This may open the model up to further attack, and may also constitute a breach of privacy
- **model extraction**: probing the model to infer how it works in order to learn how to compromise it.

The art360 toolbox provides Python functionality to evaluate and defend models against the above threats. This tool can be plugged into various machine-learning frameworks and data modalities (such as images, tables, audio and video). An alternative offering is the MIT Responsible AI Toolbox[103], which more narrowly examines Torch models using the adversarial perturbations technique (that would primarily test evasion attacks).

System developers and system owners should consult with legal teams before proceeding with some of these protections because they may implicate the company in greater risk by exposing or manipulating company data or models.

It is also important to recognise that the people interacting with the system may not always have good intentions. When a model is trained on operational data, there is a risk that it may learn undesirable behaviours from the real (but inappropriate) way that some of the users choose to interact with the system. For example, a chatbot may learn offensive word choices from its previous user interactions.[104] On a related note, online communities of enthusiasts and researchers have been demonstrating the effectiveness of "jailbreak" prompts, inputs that confuse a conversational AI such as ChatGPT or Bing AI into disregarding certain ethical or operational constraints imposed by its designers.[105]

Monitoring strategies for model and training data integrity (see Section 6) may sometimes provide early detection when a system is being attacked. Tools such as SageMaker Model Monitor[106] can alert users when there are deviations in model performance or changes in feature attribution that could be symptomatic of attacks. When monitoring anomalous behaviours, it is also important to ensure that adequate record-keeping is in place to ensure problems can be diagnosed, as discussed in Section 7. A potential gap in the resources here is readily available tools or models designed to detect and defend against black-box attack patterns, although there is active research into the topic.[107] [108]

5.3 Minimise the collection of personal information

DIRECTLY APPLICABLE TO: system owners, development team

When building an AI system it is important for the system owners to carefully consider (and justify) the data records and data attributes that are employed by the model. This relates to the principle of 'data minimisation' introduced by the EU General Data Protection Regulation (GDPR), which states that the use of personal information shall be limited to what is directly relevant and necessary to accomplish a specified purpose.[109] Limiting a system to only use the most relevant data records and attributes mitigates the risk of harm if the confidentiality of that dataset is compromised.

There also exist various tools for the development team. If a model has already been trained on a rich dataset, AI Privacy 360 provides a tool to trim down the input feature set in a performance-degradation aware manner, by either suppressing features entirely, or replacing them with less precise values (such as age ranges). Removing attributes requires the owner to forgo some performance, but the tool helps the user to balance privacy and performance deliberately.

It is important to note that removing protected attributes from a system's data may pose challenges when it comes to measuring or mitigating unfair decisions. However, we point out that it may be feasible to collect and use a supporting dataset that has appropriate permissions and compensation to the participants specifically for the purpose of evaluating fairness measures, even if the model itself does not ingest protected attributes.

Anonymising details such as names and addresses in a dataset can reduce the risk of having individuals in the records identified. This can be done manually for tabular data (by deleting columns), but this can be arduous for non-tabular data. For an algorithmic approach, Microsoft Presidio[110] offers a range of smart heuristics to automatically detect and scrub personally-identifiable information (PII) from text and image data.

5.4 Consider using privacy preserving models

DIRECTLY APPLICABLE TO: development team

A system can be designed in a way that makes it difficult to extract individual-level records from the model even if a bad actor can probe the model or has detailed knowledge of its parameters. This reduces the risk of privacy exposure.

A simple design approach (which may be applicable in limited settings) is to train the model on aggregated or synthetic data. System designers may also draw upon the algorithmic approach of differential privacy. Differential privacy is a technique to make it difficult to back out individual-level records from a model. This is achieved by adding noise, which typically comes at a cost to model performance. Toolkits with differential privacy functionality include AIP360, Microsoft Smartnoise[111], and Google Tensorflow Privacy[112].

Machine learning systems can also be designed so that an 'honest but curious' developer or collaborator does not glean private information from the data even while actively developing the system's models. This concern also applies to AI services, where the organisation may want to avoid sending input data to a third party for processing.

Homomorphic encryption applies analytics to encrypted data without decrypting it first, which may be appropriate if the dataset is highly sensitive. Tools with this functionality include Microsoft SEAL[113] and AI Privacy 360. However, the range of analyses that can be computed on encrypted data may limit the choice of models and algorithms.

If multiple parties are cooperating to develop models it is not necessary to explicitly share the data in order to benefit from each other's information. Developers might instead take a federated learning approach to share model information updates. Supporting tools in this space include Google Tensorflow Federated[114], Microsoft Azure Confidential Computing[115] (multi-party analytics) and Crypten's multi-party compute tools[116]. However, it should be noted that while this approach circumvents the need for raw data to be shared, it is still possible for a bad actor to extract private information from their counterpart's updates.

RESPONSIBLE AI PRINCIPLE

Reliability and safety

Throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose.

Curate datasets

Ensure high quality data to promote high quality model outputs.

DIRECTLY APPLICABLE TO:
system owners, development team

Conduct pilot studies

Test the assumptions of the system at a limited scale to reduce exposure to unforeseen impacts.

DIRECTLY APPLICABLE TO:
system owners, development team

Monitor and evaluate continuously

Oversee performance against both business and ethical objectives to ensure the system is operating as intended.

DIRECTLY APPLICABLE TO:
system owners, development team

6 Reliability and safety

Throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose.

The reliability and safety principle calls for businesses to ensure that the operation of a given AI system aligns with its intended purpose throughout its lifecycle, under both normal and unexpected conditions. Reliability relates to the ability of an AI system to consistently perform its intended function without unexpected or unacceptable errors or failures. Safety, on the other hand, refers to the ability of an AI system to operate without posing an unexpected or unacceptable threat to the well-being of people, society or the environment. Many engineering and data science practices employ rigour to improve the reliability and safety of AI systems. Here we focus on a limited subset particularly important for Responsible AI:

- **curate datasets** to train and validate models on accurate, representative data
- **conduct pilot studies** to evaluate an AI system in a carefully controlled environment to discover problems, iterate and scale
- **monitor and evaluate continuously** the system's data, models and performance to ensure AI systems are operating safely and reliably.

These practices are essential for controlling unintended negative impacts and ensuring that intended positive impacts are realised.

6.1 Curate datasets

DIRECTLY APPLICABLE TO: system owners, development team

The quality of the model's outputs is driven by the quality of its data. As such, data curation forms the backbone of reliability and safety in any AI model. Data curation encompasses the processes associated with the creation, management, use and maintenance of datasets. These datasets are used to inform design decisions, train models, validate the expected performance of a system and ultimately to determine whether a system is ready to progress to the next phases of the lifecycle.

As data curation requires deep technical knowledge, developers or data custodians are typically responsible for implementation, while system owners are responsible for overseeing decisions about data cleaning and exploration tools since those could create legal / reputational risk depending on how they are done. The following should be considered when developing a system.

- Failure to understand how and for what purpose historical data were collected, or how the historical data's context differs from the new deployment context, may violate modelling assumptions and may create legal risk. Tools such as Microsoft's Aether Data Documentation Template[117] suggest which metadata should be recorded to enable informed decisions about whether the datasets are fit for the envisioned purpose.
- Cleaning data and imputing missing attributes can help improve a model's performance, although both actions typically require assumptions that need to be validated and justified. OpenRefine[118] is an open source data tool to transform data and ensure that it is cleanly structured. Various cloud based platforms bundle data management tools such as SageMaker Data Wrangler (a visual interface for preparing data).[119] IBM InfoSphere QualityStage offers capabilities to cleanse and manage data (with more than 200 built in data quality detection rules).[120]
- Model evaluation requires data that is representative of the deployment use-case. This can be challenging when AI systems are typically developed using data produced by an operational system, rather than an ideal (random) sample. Open source data exploration tools such as Pandas Profiling[121] and DataPrep[122] can help data scientists understand the characteristics of the cohorts represented in the validation data through a coding-free (albeit technical) web interface. Deviations in data distribution post-deployment should be cause for concern, and can be monitored (see below).
- Datasets can be evaluated (and modified) to address fairness objectives. See Section 4 for further details.
- When using powerful and flexible models such as deep neural networks, it can be more effective for the development team to allocate resources to developing the quantity and quality of data available than to improving the algorithm itself. Here, Data-centric AI is a growing movement that shifts the engineering focus of AI systems from the design of the algorithm to the curation of the data.[123] The movement aims to establish tooling, best practices and infrastructure for managing the data used by AI systems.[124]

6.2 Conduct pilot studies

DIRECTLY APPLICABLE TO: system owners, development team

Pilot studies play an important role in the development of Responsible AI, as they are used to identify any issues or problems with a system before it is deployed more widely, and therefore can help to ensure that a system is safe, reliable and effective as it enters production.

In the context of Responsible AI, pilot studies take the form of limited-scale tests that are conducted in order to evaluate a system's performance and assess its potential impacts. Because individuals within a pilot program are still vulnerable to harm, appropriate safeguards should be put in place to protect them from risks of significant impact.

Various resources provide guidance on how to transition from pilot studies to a production scale deployment.[125] There are several important considerations to keep in mind during this process:

- Data quantity will typically increase, but it is also important to ensure the new dataset is of high quality, has representative sampling coverage, and is labelled accurately.
- Infrastructure and maintenance requirements may grow and change.
- Transparency measures in deployment may differ from the pilot (for example, trial participants may be compensated volunteers who sign a waiver).
- Model performance will need to be closely monitored (see Section 6.3 below) and any necessary revisions such as re-training or re-configuring should be made.
- Discovery of negative impacts may lead to additional objectives and metrics being introduced into the system specification.

The field of Machine Learning Model Operationalization Management (MLOps) is bridging the gap between data scientists and operations professionals in terms of applying engineering practices from software development including agile development, automated testing, release cycles and continuous integration to data science and machine learning systems in order to address some of these challenges.[126] [127]

6.3 Monitor and evaluate continuously

DIRECTLY APPLICABLE TO: system owners, development team

Conditions change over time which might invalidate the system's underlying assumptions, so monitoring an AI system's operation is an essential practice to ensure that it operates safely, reliably and effectively. It is essential that the development team measure and report metrics such as system performance (including ethical performance), and this is overseen by the system owners. As such, various best-practice guidelines have been published on the topic. [128] [129]

Monitoring can help detect a wide range of issues that impact the reliability of the system, including but not limited to:

- **dataset shift:** changes in the characteristics of the cohort using the system
- **adversarial attacks:** malicious actors attempting to exploit system vulnerabilities
- **underperforming system infrastructure:** high latency or failures due to insufficient computational resources or configuration problems.
- **data outliers:** input data points that differs significantly from what is being modelled
- **concept drift:** a change in the relationship between the input data and the feature being predicted

Many vendors are now competing to offer monitoring tools, and an exhaustive review of every product is out of scope. Some of the tools form components of a larger data science platform such as TensorBoard[130], the standard dashboard for Tensorflow projects, and SageMaker Model Monitor[131], which is part of the SageMaker platform. Others are stand-alone tools that may be deployed as either software or services.[132]

TensorBoard is notionally focused on model training and validation, but is highly extensible and also provides introspection functionality such as tracing and visualising deep neural networks for workflows developed in Python and Tensorflow. SageMaker Model Monitor is designed for in-production monitoring of models developed for the SageMaker platform, where it supports automated alerts when there are deviations in data or model quality, heuristics to monitor bias in a model's predictions or its feature attributions.

Prometheus is a popular open-source solution for logging time series data such as system metrics.[133] Metrics are logged to a server from within a system's code (with bindings provided in many languages including Python and R). This makes the approach platform agnostic, but does require developers to use other tools to compute their metrics. The logged metrics can then be ingested by extensible web-interface dashboards, for which Grafana is a popular open-source solution, providing a configurable interface with plugins, visualisation, alerts and reports, where commercial services are offered for managed cloud deployments.[134]

Key capabilities to consider when choosing a monitoring tool include:

- monitoring performance in production
  - Many tools support a range of standard metrics out of the box such as accuracy and precision, but in a Responsible AI setting it is important to also oversee ethical measures such as fairness.
  - Many tools are extensible to support custom metrics (or even offer an API to log arbitrary metrics from the model workflow)
- monitoring for problems that impact the reliability of the system (see above)
- alert delivery if a system falls below minimum baseline performance levels or a potential problem is flagged (many support real-time notification integrations with SMS or Slack for an on-call engineer, or integration to continuous integration pipelines for example)
- error traceability and explainability interfaces (see Section 7)
- experimentation, tracking metrics of different model variants to support A/B testing or feature canaries.

It is also important to consider usability factors, and their compliance implications for the business, such as:

- **how the user accesses the tool:** Many tools provide a Web UI, a REST API (implying the user might develop their own interfaces) or a console based tool (which may be more suitable for developers). Many provide graph visualisations, although this doesn't get around the need for the person responsible for oversight to have a deep understanding of what the metrics mean.
- **interoperability with development environments:** The product may be delivered as a commercial software package, open-source software, or a managed cloud service. Consider whether the tool transfers data to an external server (and whether this is acceptable).

A subtle problem that monitoring and metric evaluation tools tend to overlook is the effect of sample size. For example, fairness metrics computed over fine-grained groups may have insufficient data to draw statistically significant conclusions.

In addition to ongoing monitoring, rigorous periodic audits or algorithmic impact assessments of in-production AI systems can provide a deeper understanding of any potential issues with its performance and design, and ensure the system is up-to-date with evolving extrinsic requirements.[135]

RESPONSIBLE AI PRINCIPLE

Transparency and explainability

There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.

Make appropriate disclosures

Inform users of an AI's operation to build trust and empower them to make effective decisions.

DIRECTLY APPLICABLE TO:
system owners

Enable external scrutiny

Promote deliberate reflection within the development team and incentivise rigour through external scrutiny.

DIRECTLY APPLICABLE TO:
system owners, development team

Offer appropriate explanations

Help stakeholders understand the system's decision process to build trust and uncover problems.

DIRECTLY APPLICABLE TO:
development team

7 Transparency and explainability

There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.

Under this principle, transparency relates to sharing relevant information about the use of an AI system, while explainability relates to enabling stakeholders to understand how the outcome of an AI system was arrived at.

The level of transparency and explainability that is appropriate for a system is highly dependent on the context in which that system operates. Determining if or what to disclose, and who to address that disclosure to, requires human judgement to balance the benefits against the risks.

Promoting transparency and explainability offers a number of key benefits:

- It is fundamental to establishing trust.
- It enables external parties to identify issues with the system.
- It establishes accountability for system owners and incentivises the developers to act in a manner that can withstand public scrutiny.
- It helps set appropriate expectations around the capabilities of the system
- It may be required for compliance (e.g. with GDPR explainability requirements).

On the other hand, risks of inappropriate or excessive transparency and explainability include:

- privacy and security risks
- exposure to malicious actors manipulating the system
- exposure to audience misinterpretation
- consumer fatigue or lack of engagement.

A stakeholder's need for explainability depends on their individual preferences and their relationship to the system.

People the system makes decisions about should be aware that they are interacting with an AI system. They should also be informed about the use of their personal data, and in some contexts may expect an explanation for how the system determined their outcome, or suggestions for courses of action to change it.

People who act on the system's advice should understand the limitations of the system to prevent misuse and ensure they are equipped to justify their decisions. These might be customers who use a product or service, or staff within the organisation that use the AI system as part of their decision process.

System owners and the development team gain insight to improve their systems by diagnosing erroneous outcomes. This is relevant to many roles, such as interface designers, data scientists, developers, testers and operators. Depending on their specific role, they may need to make judgements such as whether an AI model is appropriate for their use case, or whether the model's outcomes align with its specified objectives. The system's owners will need to manage the transparency trade-offs with their organisation's other operational goals.

External reviewers may observe the system's impacts across various demographics. These reviewers may, for example, be regulators, assessors contracted by the organisation, independent auditors or academics. They may need to know the key system performance metrics so they can evaluate the success of the system against a benchmark, or may need to trace a particular outcome.

The general public, especially individuals and groups impacted by the system, should be aware of the purpose behind the design of the system and the nature of its anticipated impact (including any public harms and benefits). See Section 2 for resources on recognising a diverse set of potential impacts and impacted stakeholders.

In this report, we address three overarching categories of practices to address transparency and explainability requirements:

- **disclosing** key information about the system
- documenting key information about the system **to enable external scrutiny**
- **explaining** the system appropriately to a range of audiences.

7.1 Make appropriate disclosures

DIRECTLY APPLICABLE TO: system owners

When an AI system is engaging with people, informing users of an AI's operation builds trust and empowers them to make informed decisions. Transparency for impacted individuals could be as simple as informing the user when they are interacting with an AI system. For example, Microsoft's Responsible AI Standard[8] requires disclosure whenever the system is impersonating an interaction with a human, such as when a chatbot greets a customer.

However, even when the user is fully aware they are interacting with an AI, they might still misuse its output if they do not understand the limitations. This risk can be elevated if a generative model produces sufficiently compelling content. For example, ChatGPT is known to hallucinate incorrect facts or even references, which a user might assume to be correct if they have not read the full disclaimers.

Depending on the context and use-case it may also be appropriate for system owners to publish relevant information pertaining to the purpose and limitations of the AI system. For systems whose impacts are considered more significant, providing additional details about the system can be a step towards establishing legitimacy and social licence. Industry practitioners are increasingly taking the measure of sharing relatively accessible details of their models publicly.

Algorithmic impact assessments are conducted internally to identify potential system harms (as outlined in Section 2), but a growing number of organisations are publishing their reports or encouraging the public to participate in the drafting process. *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest*[136] provides guidance around assessment of a system with public consultation and an independent assessor. This process can be arduous, so is typically used by organisations that place a heavy emphasis on public trust and accountability such as public bodies. For example, The Office of Qualifications and Examinations Regulation (Ofqual) in the UK conducted rounds of public consultation prior to the implementation of their examination grading and assessment algorithm in 2020 (although the system was ultimately determined to be not fit for purpose).[137]

AI system registers can help to establish clear visibility over which algorithmic systems are in use. Additional metadata recorded in the register can be designed to enable meaningful transparency for particular stakeholders. For example, the New Zealand Government has published a review[138] of all their AI systems that interact with the public, disclosing the systems' purposes as well as the checks and balances that are in place to manage their use. The draft EU AI Act (2021) proposes to mandate that organisations publish inventories of their AI systems and their associated risk mitigation measures going forwards.[139] The local governments in the cities of Helsinki[140] and Amsterdam[141] have released AI system registers describing relevant information to the public such as where and how the AI systems are used, what data and algorithms are employed, how the systems impact the lives of citizens as well as contact information for the systems' development teams.

7.2 Enable external scrutiny

DIRECTLY APPLICABLE TO: system owners, development team

Documentation is essential to support transparency. Internal users may need technical documentation to inform design decisions, while external reviewers may need documentation that is accessible to their specific domain of expertise (such as legal or regulatory compliance). The process of documenting can also help the development team to engage in deliberate reflection on the impact of the design decisions they make. In addition to the impact assessments and system registers mentioned above, several approaches are becoming popular with AI practitioners.

System fact sheets are analogous to the nutrition fact sheets that appear on food products at the supermarket. They are aimed at quickly informing those outside the immediate development team, such as the general public, users, procurers and auditors, about key aspects of the system. These sheets should, at a minimum, contain information about the system's purpose, intended uses and limitations. Additional details often explore the system's input data, its performance on validation sets and details of the assumptions that underpin the model. Google[142], IBM[143], and Microsoft[144] have released their own templates, with worked examples. OpenAI published a system card to accompany the launch of GPT-4 large language model which discusses topics such as safety challenges, deployment preparations and next steps.[145]

Data factsheets summarise properties of the data used to train and evaluate the models, exploring the data's quality, representativeness and the assumptions that guided the preprocessing phase. A key example is the Aether Data Documentation Template[146], an evolution of the popular Datasheets for Datasets[147] resource.

System decision registries are a practice employed across a wide range of industries to promote accountability and transparency.[148] [149] Decision registries provide a mechanism for teams to record key choices that were made during the development of the AI system, to identify who made them and to justify the rationale behind them. The registry can be invaluable for:

- ensuring continuity of knowledge as personnel change throughout the model lifecycle
- checking that the values implicit in the system align with that of the organisation
- clearly identifying who is responsible for a decision and incentivising them to make choices that are considered acceptable within the organisation.

7.3 Offer appropriate explanations

DIRECTLY APPLICABLE TO: development team

Explanations help stakeholders understand the system's decision process to build trust in its decisions.

This does not necessarily mean that stakeholders need to be able to interpret how the model works by examining the algorithm, or its parameters (nor should they be expected to). Interpretability and explainability mean subtly different things in the context of Responsible AI. Interpretability is the degree to which a human can understand the cause of a decision, or predict the model's decision by examining the decision process.[150] [151] An inherently interpretable model is necessarily low complexity such that a person can follow its workings, such as a linear model.

Explainability, on the other hand, relates to conveying an effective mental model of the system's decision process to a stakeholder, even if they don't fully understand the internal workings. This may involve simplifying the true decision process to capture the most relevant factors and derive generalisable insights. A challenge here is to ensure the explanation is audience appropriate, as stakeholders (including end users and consumers) will possess various levels of background knowledge and have differing goals. We note that many AI model explanation software tools have been developed for data scientists, while relatively few resources are suitable for other stakeholders. As such, the development team is primarily responsible for generating explanations to the system owner's specifications.

Criteria for explainability

Interpretable Machine Learning[152] (which also discusses explainability) identifies multiple criteria for an explanation to be effective, including being contrastive (focusing on why one outcome was selected instead of another), selective (focusing on the most relevant contributing factors), consistent with the audience's understanding of the application domain, and for the explanation to generalise to similar cases (so the explanation helps the audience predict what the model will do). Additional discussion can be found in the literature.[153]

Stakeholders may need to understand what data was used, how the model used this data, or why the system has arrived at a particular decision. For some *inherently interpretable* models (low complexity, with well understood parameters), explanation is relatively straightforward, as it is possible to interpret their parameters directly by:

- examining the weights in a linear model
- displaying the rules in a decision tree
- providing the neighbours in a nearest-neighbour model.

Notably, InterpretML[154] provides a suite of inherently interpretable models for practitioners to use. In practice, however, many systems use complex models because of their higher performance potential. In this case, effective post-hoc explanations require a deliberate balance between detail (describing precisely what is going on in the model) and simplicity (limiting the complexity of the description).

Tools for explainability

Approaches are split into two camps in that they are either:

- **global explanations**, which apply broadly to the whole cohort, but aren't necessarily accurate for every decision
- **local explanations**, which accurately describe a specific decision, but don't necessarily generalise.

When deciding which is more appropriate, a practitioner might ask whether the stakeholders receiving the explanation are more interested in broadly "how" the model works (global), or specifically "why" the model has come to a particular decision (local). Most toolkits in this space offer a range of both local and global techniques.

Other tools are designed to extract insight from the internal workings of a specific complex model. Popular examples include random forest feature permutation analysis, and neural network interpretability analyses including:

- attribution approaches such as saliency maps to identify important inputs or intermediate representations, and
- feature reconstructions that reveal the learned purpose of internal representations.

Toolkits that specialise in neural network interpretations include Google's Language Interpretability Tool[155] (LIT), and Captum.[156] Both (surprisingly given the former's name) provide analyses for tabular, language and image data, or a combination of modalities, a key differentiator being that LIT wraps the dataset and model, while Captum integrates closely with a Torch workflow. For those using deep learning in cloud environments, SageMaker Clarify provides attribution methods for the SageMaker ML platform, while Google's *Explainable AI (xAI)* connects Google's ML services (*AutoML Tables*, *Bigquery ML*, *Vertex AI*) to their What-If Tool and LIT functionality.

On the other hand, model-agnostic methods can be applied to any model. These techniques require no knowledge of the underlying algorithm and its parameters. "Black box" feature importance is a ubiquitous example, attributing the degree to which particular input features are driving the model predictions. Many toolkits provide feature importance methods, including interpretML, aix360, SageMaker Clarify, Captum, with SHAP[157] being a commonly supported attribution algorithm.

A view of *how* a particular feature drives the predictions can be provided by techniques such as partial dependence plots (PDP)[158] or individual conditional expectation (ICE)[159] plots. Alternatively, an interpretable model can approximate the decision process for the purposes of explanation: LIME[160] is a popular candidate. Notably, InterpretML and aix360, the two toolkits aimed at a Python / Scikit learn workflow, provide LIME analysis, while aix360, LIT, SageMaker Clarify and the What-If Tool provide partial dependence plots.

Example-based approaches, as opposed to feature-based approaches, work by presenting the audience with similar, contrasting or representative prototype data. Contrastive approaches may highlight minimally sufficient features and critically absent features. Counterfactual approaches try to explain why the actual outcome was chosen instead of an alternative, by creating a slightly modified version of the input which results in a different outcome, allowing "what if" type questions to be answered. Data explorers and visualisation tools are a digestible approach to using example-based explanations, allowing the user to examine both data and model outcomes over a cohort. The two main offerings in this space are the What-If Tool[161] and Error Analysis[162]. A key differentiator here is that the What-If Tool is easily integrated into a Tensorflow workflow but operates on a static dataset (which may contain model predictions or scores), while Error Analysis can also be hooked up to any Python model for active querying – meaning it can create synthetic counterfactuals, or conduct black-box feature importance analysis.

7.4 Resource gaps and considerations

Although there are many resources available for addressing transparency and generating explanations of algorithmic decisions, some gaps remain in their coverage.

- Explainable AI tools are aimed almost exclusively at developers and as such demand a high degree of technical knowledge to operate and interpret their outputs.
- Explainable AI algorithms are not the only solution to improving system explainability. Explainability can also be addressed by designing effective explanation interfaces and understanding the psychology of explanations. The role of explanations may be informed by the level of trust users have in the system.

- Objective measures of explanation quality are lacking. AI Explainability 360 outlines two proxies that can be used: faithfulness and monotonicity.[163] Alternatively, effectiveness might best be measured with audience-in-the-loop experiments such as measuring the audience's ability to predict model behaviour.[164]
- We caution that feature importance explanations are commonly misleading:
  - Importance means a model prediction is strongly influenced by the feature. Domain knowledge is required to decide why this might work, and whether it is a sensible feature to base a prediction on. However, machine learning often uses a rich combination of co-dependent features. If a seemingly "important" feature is removed, the importance often simply hops onto other features.
  - Intervening on important model features may not have the intended real-world outcome. It may be that the feature drives the outcome, but it is equally possible it is driven by the outcome, or both are caused by common extrinsic factors.
- Generating personalised explanations that advise a person how to act (algorithmic recourse) is an open problem.[165] Rather than explaining with the goal of audience understanding, recourse explanations need to be feasible and inform actions that have a beneficial outcome for the recipient. This places strong requirements on the explanation to:
  - use audience-appropriate language or diagrams
  - suggest actionable changes (for example, suggesting that someone establishes a credit history, rather than suggesting that someone change their age or ethnicity)
  - factor in the side-effects of advice (for example advising someone to move to a higher paying job to get a loan may increase their salary, but reset their tenure).
- In many settings, such as education, the use of generative models may be prohibited or restricted. The publishers of the tool may be able to promote user compliance by introducing content watermarking, or providing tools that detect generated outputs,[166] although such technologies are still emerging and users could potentially subvert them by making modifications to the content.

RESPONSIBLE AI PRINCIPLE

Contestability

When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.

Understand contestability obligations

Determine proportionate means for affected stakeholders to voice their objections.

DIRECTLY APPLICABLE TO:
senior directors, system owners

Establish human review of contested decisions

Allow human decision makers to provide context and compassion that the system may lack.

DIRECTLY APPLICABLE TO:
system owners

Support impacted individuals

Provide the means for an impacted individual to mount a successful contest.

DIRECTLY APPLICABLE TO:
senior directors, system owners, development team

8 Contestability

When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.

Providing efficient and accessible mechanisms to contest an AI system's decisions can be important or even essential, especially when these decisions significantly impact vulnerable individuals and communities. Contestability helps to establish consumer trust and keep businesses accountable. Some publications even suggest it should be a fundamental democratic right.[167] [168]

This section points to resources businesses can use to address contestability and highlights current gaps in their coverage. We examine these grouped into four relevant practices:

- understanding **contestability obligations**, as not every use of AI warrants a contestability process
- establishing **human review** of contested decisions
- supporting **impacted individuals** to mount a successful contest.

8.1 Understand contestability obligations

DIRECTLY APPLICABLE TO: senior directors, system owners

Understanding contestability obligations is prerequisite to establishing proportionate means for impacted individuals to challenge decisions that affect them. One criterion for determining whether contestability is appropriate is the potential for *significant impact*. Australia's Artificial Intelligence Ethics Framework states that contestability is required when an AI system significantly impacts a person, community, group or environment.[1] Similarly, the GDPR's Article 22(3) states that it is a person's legal right to contest a decision they are significantly impacted by.[169] Deciding whether the impact is significant is a highly contextual decision that needs to be made by those accountable for the operation of the AI system, who should be prepared to justify the decision. The system owner must collaborate with specialists in the relevant domain and legal professionals to comprehensively ascertain the requirements for contestability.

8.2 Establish human review of contested decisions

DIRECTLY APPLICABLE TO: system owners

Having a human decision maker review contested decisions allows them to bring in context and compassion that the system may lack. Ultimately, the decision to uphold or overturn an algorithmic decision must fall to a responsible human decision maker.[170] [171] As such, the process for contesting algorithmic decisions is often modelled on the way people already review decisions in the organisation.

Businesses should proactively identify potential areas of contest and clarify who is responsible for addressing them at each step of the review process, at every stage of the AI system lifecycle.[172] Doing so will ensure effective review of contested outcomes as they arise.

Businesses may also have existing dispute resolution channels that they can leverage, provided that they incorporate suitable human oversight as discussed above. For example, some financial and telecommunication services have internal dispute processes in place to address customer concerns.[173]

8.3 Support impacted individuals

DIRECTLY APPLICABLE TO: senior directors, system owners, development team

AI systems often employ complex algorithms with opaque decision processes. It is nonetheless essential to provide impacted individuals with an adequate understanding of how the system decided their outcome and what data the decision was based on so that they have grounds to contest.

We discuss three key considerations for businesses when providing impacted people with a basis to contest:

- providing clear explanations and justifications for the decisions made (see Section 7)
- establishing recourse and redress mechanisms
- adopting preemptive contestability so people can challenge a decision before it is enacted

Recourse and redress mechanisms

Correcting erroneous decisions or remedying their impacts reduces harms the system may otherwise cause.[174] Two concepts are relevant to the outcomes of contestability:

- ***recourse:*** the capacity of an impacted individual to alter the decision made by an AI system by utilising mechanisms that initiate a thorough review process
- ***redress:*** taking action to remedy or set right harms resulting from an AI decision that has undesirably or unfairly impacted a person or community.

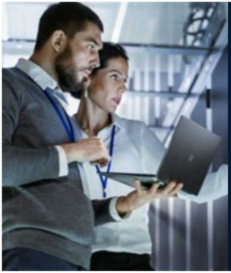
For further reading, organisations might look to foreign standards that have recognised the need for recourse and redress when decisions are made by AI systems or are based on human data. The European Commission's Ethics Guidelines for Trustworthy AI explains the importance of adopting redress mechanisms to ensure trust, and that accountability frameworks should include review and redress mechanisms.[16] The UK's Data Protection Act 2018 Article 46 demands the 'right to rectification', where the company controlling the data must rectify incomplete or inaccurate data upon request by the data subject (subject to conditions).[175]

Preemptive contestability

When errors are anticipated to happen frequently, preemptive contestability allows erroneous decisions to be challenged before they generate impacts. This can help manage harms, and build trust in the system. One way to do this is to show the affected person their outcome before it is acted upon.[3] This is complementary to recourse strategies that control harms that have already occurred.

RESPONSIBLE AI PRINCIPLE

Accountability



Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

Raise awareness in Responsible AI

Ensure that individuals are equipped to make ethical decisions when designing and deploying AI systems.

DIRECTLY APPLICABLE TO: senior directors, system owners, development team

Establish roles and responsibilities

Be clear about who is accountable for different aspects of the AI system's operation and impacts.

DIRECTLY APPLICABLE TO: senior directors, system owners, development team

Conduct independent external audits

Seek unbiased evaluation of the system's performance and its delivery on its intended purpose to identify potential issues.

DIRECTLY APPLICABLE TO: senior directors, system owners, development team

Create positive incentives

Drive and reinforce responsible AI practices by explicitly motivating ethical behaviour.

DIRECTLY APPLICABLE TO: senior directors

9 Accountability

Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

Accountability in a system can be thought of as an assurance mechanism that promotes alignment between the actions of individuals and the purpose of the system. If the purpose of a system is to give effect to all the principles previously listed, then accountability functions as a "meta-principle" that can promote responsible AI by furthering each of these principles.

Accountability is the concept of being answerable for one's actions and decisions. It is the idea that individuals and organisations should be held responsible for the consequences of their actions, and should be able to provide a clear and satisfactory explanation for these (and potentially be subject to penalties for wrong or inappropriate conduct). More precisely, accountability is a relationship involving:[176] [177]

- an accountable party
- a clear scope of responsibility for the accountable party
- a forum to whom the accountable party is answerable within the scope of its responsibility, where answerability amounts to:
  - the capacity of the accountable party to explain and justify its decisions to the forum
  - the right of the forum to request from the accountable party explanations and justifications for its decisions to the forum
- a potential right from the forum to impose on the accountable party penalties or corrective measures.

This implies that appropriate accountability depends on the appropriate assignment of roles and responsibilities and on knowledge, tools, resources and transparency required for effective answerability.

When an organisation has a culture of accountability, as well as adequate systems giving effect to that culture, people become appropriately resourced, motivated and incentivised to act in the best interests of the organisation, reducing the risk of unintended consequences. When this assurance includes the use of AI systems, it helps build trust and confidence in the use of AI by the organisation as people recognise such systems as trustworthy.

Below are some key practices that can improve accountability in the use of AI systems:

- From a people and culture perspective, **education** on the responsible use of AI can build awareness of the relationship between individual decisions and actions by the workforce – from detailed technical decisions to broader strategic decisions – and the human impact caused by the organisation's AI systems. This can ignite a stronger sense of responsibility and duty towards the impacted customers and other people affected.
- From a process perspective, establishing appropriate **roles and responsibilities** and ensuring the roles are properly orchestrated can improve individual and system-level accountability.
- From a broader governance and risk perspective, practices such as **independent external audits** can be effective at detecting blind spots and unacknowledged issues.
- From an organisational culture perspective, providing **positive incentives** for system owners, developers and stakeholders promotes the adoption of responsible AI practices.

9.1 Raise awareness and knowledge of Responsible AI

DIRECTLY APPLICABLE TO: senior directors, system owners, development team

By providing employees with the knowledge and skills to use AI in a responsible and ethical manner, an organisation can help ensure that the decision-making processes implemented through its AI systems are compliant with the law and adequately aligned with the organisation's values. This can help to improve the transparency and accountability of the organisation's AI systems and reduce the risk of ethical or legal issues arising from their use.

An organisation can provide training on the responsible use of AI to its workforce through a variety of different methods, including training programs for different groups of employees such as data scientists, managers, and directors. Some examples of these methods include:

- **workshops and seminars:** Training sessions on the responsible use of AI tailored to the specific needs and roles of different groups of employees. For example, data scientists can receive training on how to develop, configure, operate and deploy AI systems with a perspective centred on their human impacts, while managers can receive training on how to resource their teams, oversee and monitor their technical work with the aim of aligning it with the instructions from senior directors, and senior directors on the broader risk and governance considerations of using AI systems in an organisation.
- **online courses and training materials:** Access to online courses and training materials on the responsible use of AI. This can include introductory topics such as a general introduction to responsible AI and a technical introduction to responsible AI, as well as technically oriented courses such as foundations of decision theory, machine learning and automated decision-making, algorithmic fairness, explainability in AI, and AI risk and governance courses for senior directors. Many resources are available for RAI training.[178] [179] [180] [181] [182] [183]
- **hands-on training:** Training to help employees develop practical skills and experience in the responsible use of AI. For example, data scientists can be given access to AI development platforms and tools, and can be provided with guidance and support as they work on projects that involve the development and deployment of AI systems.
- **external experts and speakers:** The organisation can bring in external experts and speakers to provide training and insights on the responsible use of AI. This can include experts in AI ethics, law, and policy, as well as practitioners who have experience in developing and deploying AI systems in a responsible manner.

Providing training on the responsible use of AI to the workforce is a foundational practice to promote accountability for the use of AI systems within an organisation. It paves the way for the introduction of other accountability-promoting practices as the workforce is made more conscious of the imperative to manage AI systems responsibly.

9.2 Establish roles and responsibilities

DIRECTLY APPLICABLE TO: senior directors, system owners, development team

Clarifying the roles and responsibilities of different individuals and stakeholders within an organisation can help improve its entire system of accountability when it comes to actions and decisions made by AI systems.

In the report *De-risking Automated Decisions: Practical Guidance for AI Governance*, Gradient Institute addresses how different parties within an organisation can have the scope of their roles and responsibilities adapted to better serve the purpose of promoting effective organisational accountability in the use of AI.[3] The following are examples of a few responsibilities that should be expected from some of the key stakeholders involved in the governance of an AI system:

- **board of directors:** The board is ultimately accountable for what the organisation does, including the operation of its AI systems, and as such it has the responsibility to seek appropriate education on the risks of using AI systems and how to put in place effective governance mechanisms to control and monitor the use of AI systems in the organisation.
- **business/system owners and integrators:** These parties are in charge of running the AI systems, i.e. defining the business and other objectives that the AI systems should be configured to optimise as well as ensuring that the implementation is fit for purpose and delivers on the objectives. As such, they are responsible for deciding how to balance potentially conflicting objectives (such as revenue, efficiency and ethical restraints) in the best interests of the organisation, according to direction from the board and senior executives. They are also responsible for ensuring that their teams are appropriately resourced, instructed and overseen so as to accurately translate those higher level decisions to the technical design and implementation of the AI system. In addition they need to ensure that they have appropriate visibility of the technical design decisions so they can be capable of explaining and justifying the business and ethical relevance of those to the upper echelons of the organisation.
- **developers and data scientists:** These technical staff are primarily responsible for understanding the relationship between design and implementation decisions and the actual decisions the system produces. They should have an adequate understanding of the business domain in which they are operating, and be able to communicate technical design decisions to non-technical audiences (in particular the business/system owners). They also need to be capable of representing non-mathematical objectives (such as "ethical" objectives articulated by business owners) in mathematical form to be interpreted by the AI system; when doing so, they need to be able to explain the limitations and risks associated with such approximations so the business owner can provide feedback about the appropriateness of any particular assumptions. The developers should also be in charge of identifying unintended behaviour of the AI systems and their causes, as well as of implementing effective corrective and mitigation strategies.
- **AI governance committee:** An AI governance committee can be put in place and one of its responsibilities should be to assess the degree to which a system's design and operation aligns with its stated objectives, and the values and priorities of the organisation (regardless of whether the system is developed in-house or procured). The committee needs to understand the limitations of the system as implemented and its potential unintended impacts.

A clear definition of roles and responsibilities all the way from the boardroom down to the design and implementation of the AI system by developers and data scientists is crucial to assemble an effective system of accountability for an AI system's decisions. Specific advice on the nature of such responsibilities needs to be cognisant of the need to establish clear objectives (business, ethical, and others) and effective means to accurately translate them to the implementation of the AI system.

9.3 Conduct independent external audits

DIRECTLY APPLICABLE TO: senior directors, system owners, development team

An external, independent audit of an AI system is a comprehensive evaluation of the performance, accuracy, and effectiveness of the system with respect to delivering on its intended purpose. Audits help with accountability because they can detect potential issues with the AI system and provide recommendations for improvement. Senior directors may incorporate independent audits as part of their AI system governance. System owners (supported by their development teams) can then develop relevant documentation and materials to conduct them.

Internationally, multiple bodies and organisations have published guidelines or standards recommending independent external audits of AI systems, including:

- The European Union[184] [185]
- The Australian Human Rights Commission (AHRC)[186]
- The United States' National Institute of Standards and Technology (NIST)[187]
- The International Organization for Standardization (ISO)[188]

When an organisation commissions an external audit of an AI system, it may expect the auditor to address a range of aspects considering the system, such as:

- **identifying the goals and objectives of the AI system:** This involves clearly identifying the purpose and intended use of the AI system, as well as the specific metrics and criteria that will be used to evaluate its performance.
- **analysing the data that the system uses:** Examining the relevance and quality of the data the AI system uses for the purpose of achieving the goals.
- **evaluating and testing the process that generates the outputs of the AI system:** Providing a technical analysis of how the algorithms and models in the system use the data to produce the systems' outputs. This is a critical step that seeks to establish the relationship between data, algorithms and models on one side and the actual final decisions the AI system produces on the other side.
- **monitoring the system:** This can involve regularly reviewing the system's output and comparing it to expected results to identify any potential issues.
- **reporting:** Once the audit is complete, detail the findings and recommendations for improving the AI system. The report is to be shared with stakeholders, such as the AI system's developers, system owner, users, and senior directors, to help inform future development and use of the system.

A variant on AI system audits is an algorithmic impact assessment (AIA), which involves, in addition to the elements mentioned above, a more in-depth analysis of the relationship between the decisions made by the AI system and the actual human impacts they create. Resources for conducting an assessment are discussed in Section 2.

9.4 Create positive incentives

DIRECTLY APPLICABLE TO: senior directors

Accountability is often framed around penalties, but positive incentives also play an important role in driving and reinforcing responsible AI practices within an organisation.

An RAI approach requires system owners to deliberately balance ethical objectives (such as fairness) against competing business objectives (such as revenue). It would be problematic for an organisation if performance evaluations and KPIs focus solely on business objectives and overlook the ethical value of people's actions.[3] Considerations for positive reinforcement include:

- recognising ethical objectives and initiatives to promote ethical outcomes in employee performance evaluations
- allocating the responsibility for setting and balancing both business and ethical objectives to the same owners so that they do not have a vested interest in tipping the balance one way or another (they may be held accountable by their organisation's AI governance committee of, or board of directors for example).

10 Conclusion

The purpose of this report is to help small and medium enterprises, and businesses in general, to start putting the Australian AI Ethics Principles into practice in their organisations.

For each of the eight principles,

- **human, societal and environmental wellbeing**
- **human-centred values**
- **fairness**
- **privacy protection and security**
- **reliability and safety**
- **transparency and explainability**
- **contestability**
- **accountability,**

the report focuses on three tasks:

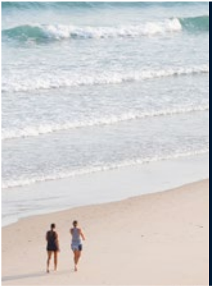
- **suggesting some key practices that a business can cultivate in order to promote the principle**
- **pointing to resources to help conduct the selected practices**
- **when there are gaps in existing resources, suggesting alternative courses of action.**

AI technology is evolving at an astonishing pace. Readers of this report should keep in mind that choosing an appropriate tool to support a certain practice will continue to be a hard problem to solve as new tools and resources keep on emerging. This suggests that it is advisable that organisations adopt a proactive attitude to keeping themselves abreast of the latest developments in responsible AI resources and tools.

Even though the tools will keep on evolving at a fast pace, it should be noted that many practices are likely to stay relevant and new practices will emerge. This suggests that it is advisable that organisations invest in developing their culture and governance processes so as to eventually elevate Responsible AI to a level of standard routine – in a way that is agnostic to the particular choice of tools or resources required for execution. The need to retire practices or create new ones will eventually arise, but this should not distract organisations from the task of instituting and developing practices that are today known to be effective and are likely to continue to be for the foreseeable future.

Responsible AI – selected practices

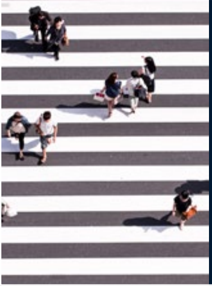
Human, societal and environmental wellbeing



Throughout their lifecycle, AI systems should benefit individuals, society and the environment.

Elicit potential impacts	Assess impacts	Set ethical objectives
👤 </>	👤 </>	👤 </>

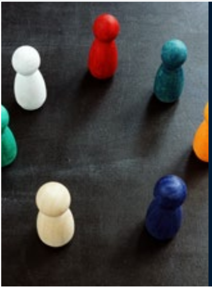
Human-centred values



Throughout their lifecycle, AI systems should respect human rights, diversity, and the autonomy of individuals.

Design for human autonomy	Achieve outcomes ethically	Incorporate diversity
</>	👤	☷ 👤 </>

Fairness



Throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.

Contextualise fairness	Measure fairness	Mitigate unfair impacts
☷ 👤	👤 </>	👤 </>

Privacy protection and security

Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.

Consult privacy and security experts	Guard against attacks	Minimise the collection of personal information	Consider using privacy preserving models
☷ 👤	</>	👤 </>	</>

ICON DEFINITIONS – This document is written for senior directors, system owners and system developers. These icons indicate which of these three roles each practice is directly applicable to.

👤 **SYSTEM OWNER** – the person (or persons) responsible for defining business and other objectives of the system (in line with the board of directors' strategy), as well as for ensuring that the system implementation is fit for purpose and delivers on those objectives.

Reliability and safety

Throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose.

Curate datasets	Conduct pilot studies	Monitor and evaluate continuously
🔍 </>	🔍 </>	🔍 </>

Transparency and explainability

There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.

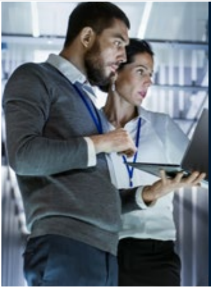
Make appropriate disclosures	Enable external scrutiny	Offer appropriate explanations
🔍	🔍 </>	</>

Contestability

When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.

Understand contestability obligations	Establish human review of contested decisions	Support impacted individuals
👤 🔍	🔍	👤 🔍 </>

Accountability



Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

Raise awareness in Respon-sible AI	Establish roles and respons-ibilities	Conduct independ-ent external audits	Create positive incentives
👤 🔍 </>	👤 🔍 </>	👤 🔍 </>	👤

👤 **SENIOR DIRECTORS – concerned with setting strategic goals and ensuring the organisation's activities are aligned with those goals.**

</> **DEVELOPMENT TEAM – responsible for designing and implementing the AI system to meet the objectives and requirements specified by the system owner.**

References

1 Department of Industry, Science and Resources (n.d.). *Australia's AI Ethics Principles.* Australian Government. https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles

2 Fjeld, J. *et al.* (2020) "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI," *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.3518482.

3 Caetano, T., Davis, J., Dolman, C., O'Callaghan, S., & Weatherall, K. (2022). *De-Risking Automated Decisions: Practical Guidance for AI Governance.* Gradient Institute supported by Minderoo Foundation. https://www.gradientinstitute.org/assets/gradient_minderoo_report.pdf

4 UTS Open (2022). *Short Course - Ethical AI: from Principles to Practice*. UTS. https://open.uts.edu.au/uts-open/study-area/Technology/AI--ML/ethical-ai-from-principles-to-practice/

5 People + AI Research Team (n.d.). *Feedback + Control*. People + AI Guidebook. https://pair.withgoogle.com/chapter/feedback-controls/

6 High-Level Expert Group on AI (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

7 Veritas (2022). *Veritas Document 3A: FEAT Fairness Principles Assessment Methodology.* Veritas. https://www.mas.gov.sg/-/media/MAS-Media-Library/news/media-releases/2022/Veritas-Document-3A---FEAT-Fairness-Principles-Assessment-Methodology.pdf

8 Microsoft (2022). *Microsoft Responsible AI Standard, v2*. Microsoft. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

9 Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector.* The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

10 AI Ethics Lab (2021). TOOLBOX: Dynamics of AI Principles. *AI Ethics Lab*. https://aiethicslab.com/big-picture/

11 Tech Policy Lab (2022). *Diverse Voices: A How-To Guide for Facilitating Inclusiveness in Tech Policy*. University of Washington. https://techpolicylab.uw.edu/project/diverse-voices/

12 ADAPT Centre (2017). *The Online Ethics Canvas.* The ADAPT Centre for Digital Content Technology. https://ethicscanvas.org/

13 Microsoft Azure (2022) *Judgment Call*. Microsoft Learn. https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/judgmentcall

14 O'Neil, C., & Gunn, H. (2020). Near-term Artificial Intelligence and the ethical matrix. *Ethics of Artificial Intelligence*, 237–270. https://doi.org/10.1093/oso/9780190905033.003.0009

15 Omidyar Network (2018). *Ethical OS Toolkit*. Omidyar Network with Institute for the Future. https://ethicalos.org/

16 High-Level Expert Group on AI (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

17 Microsoft Azure team (2023). *Azure Application Architecture Guide - Types of harm*. https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/type-of-harm

18 Government of Canada (2022). *Algorithmic Impact Assessment tool*. Government of Canada. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

19 Microsoft AI Team (2022). *Responsible AI*. Microsoft. https://www.microsoft.com/en-us/ai/responsible-ai

20 Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). *Algorithmic Impact Assessments.* AI Now Institute. https://ainowinstitute.org/aiareport2018.pdf

21 ISO. (n.d.). *ISO/IEC AWI 42005 Information technology - Artificial intelligence - AI system impact assessment.* ISO. https://www.iso.org/standard/44545.html

22 Bird, S. (2022). *Responsible AI investments and safeguards for facial recognition.* Microsoft Azure. https://azure.microsoft.com/en-us/blog/responsible-ai-investments-and-safeguards-for-facial-recognition/

23 Li, D., Hasanaj & E., Li, S. (2020). *3 - Baselines*. Machine Learning, Carnegie Mellon University. https://blog.ml.cmu.edu/2020/08/31/3-baselines/

24 Ronaghan, S. (2019). *Statistical Tests for Comparing Machine Learning and Baseline Performance.* Towards Data Science. Medium. https://towardsdatascience.com/statistical-tests-for-comparing-machine-learning-and-baseline-performance-4dfc9402e46f

25 Ameisen, E. (2018). *Always start with a stupid model, no exceptions.* Insight Data Science. Medium. https://blog.insightdatascience.com/always-start-with-a-stupid-model-no-exceptions-3a22314b9aaa

26 Tenini, J. (2019). How to do Cost-Sensitive Learning. Red Ventures Data Science & Engineering. Medium. https://medium.com/rv-data/how-to-do-cost-sensitive-learning-61848bf4f5e7

27 Howard, J., Zwemer, M., & Loukides, M. (2012). *Designing great data products.* O'Reilly. The Drivetrain Approach: A four-step process for building data products

28 Ma, P., Du, T., & Matusik, W. (2020). Efficient Continuous Pareto Exploration in Multi-Task Learning. *arXiv*. https://doi.org/10.48550/arXiv.2006.16434

29 Candido, T. (2021). *MLDrops - Optimise one evaluation metric and satisfy all other metrics.* Towards Data Science. Medium. https://towardsdatascience.com/mldrops-optimize-one-evaluation-metric-and-satisfy-all-other-metrics-367457d3c32d

30 Emmerich, M. T., & Deutz, A. H. (2018). A tutorial on multiobjective optimization: Fundamentals and Evolutionary Methods. *Natural Computing*, *17*(3), 585–609. https://doi.org/10.1007/s11047-018-9685-y

31 McCalman, L. (2022). *AI Impact Control Panel.* Gradient Institute & Minderoo Foundation. Medium. https://medium.com/gradient-institute/ai-impact-control-panel-8f2316505a1f

32 Miettinen, K., Eskelinen, P., Ruiz, F., & Luque, M. (2010). Nautilus method: An interactive technique in multiobjective optimization based on the nadir point. *European Journal of Operational Research*, *206*(2), 426–434. https://doi.org/10.1016/j.ejor.2010.02.041

33 Brignull, H. (2022). *Deceptive Design.* Deceptive Design. https://www.deceptive.design/

34 Blakkarly, J. (2022). Online consumers harmed by 'dark patterns' in web design. Choice. https://www.choice.com.au/consumers-and-data/data-collection-and-use/how-your-data-is-used/articles/dark-patterns-cprc

35 Cannon, J. C. (2015). *The Problem with Online Opt Out.* LinkedIn Pulse. https://www.linkedin.com/pulse/problem-online-opt-out-jc-cannon

36 Google (n.d.). *My Ad Centre*. Google. https://myadcenter.google.com/

37 König, P. D. (2022). Challenges in enabling user control over algorithm-based services. *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01395-1

38 McQuinn, A. (2017). *The Economics of "Opt-Out" Versus "Opt-In" Privacy Rules.* Information Technology & Innovation Foundation (ITIF). https://itif.org/publications/2017/10/06/economics-opt-out-versus-opt-in-privacy-rules/

39 Association for Psychological Science (APS) (2013). *The Opt-Out Option.* APS. https://www.psychologicalscience.org/news/minds-business/the-opt-out-option.html

40 Hyde, M. K., Masser, B. M., Edwards, R. A., & Ferguson, E. (2021). Australian Perspectives on Opt-In and Opt-Out Consent Systems for Deceased Organ Donation. *Progress in Transplantation*. https://doi.org/10.1177/15269248211046023

41 Blesch, W. (2022). *Consent Opt-In/Opt-Out Best Practices and Compliance*. TermsFeed. https://www.termsfeed.com/blog/consent-opt-in-out-best-practices-compliance/

42 Dearie, K. (2021). *Opt In vs. Opt Out.* Termly. https://termly.io/resources/articles/opt-in-vs-opt-out/

43 CookieYes. (2022). *Opt-in Vs Opt-out: What they are and How to Implement Each.* CookieYes. https://www.cookieyes.com/blog/opt-in-opt-out/

44 Office of the Australian Information Commissioner (2018). *Australian entities and the EU General Data Protection Regulation (GDPR)*. Australian Government. https://www.oaic.gov.au/privacy/guidance-and-advice/australian-entities-and-the-eu-general-data-protection-regulation

45 Australian Human Rights Commission (2020). *Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias.* Australian Human Rights Commission. https://humanrights.gov.au/our-work/rights-and-freedoms/publications/using-artificial-intelligence-make-decisions-addressing

46 Westpac Group (2021). *Westpac Group Code of Conduct*. Westpac. https://www.westpac.com.au/docs/pdf/aw/code-of-conduct.pdf

47 OECD framework for the classification of AI Systems. (2022). *OECD Digital Economy Papers*. https://doi.org/10.1787/cb6d9eca-en

48 Somepalli, G. et. al. (2022) *Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models*

49 Office of Best Practice Regulation (2020). *Best Practice Consultation.* Department of the Prime Minister and Cabinet. Australian Government. https://www.pmc.gov.au/sites/default/files/publications/best-practice-consultation_0.pdf

50 Caron, M. S., & Gupta, A. (2020). The Social Contract for AI. *arXiv*. https://doi.org/10.48550/arXiv.2006.08140

51 AWIS: AI World Society (2020). *Social Contract for the AI Age.*MIT's Sociotechnical Systems Research Center. https://ssrc.mit.edu/wp-content/uploads/2020/10/Social-Contract-for-the-AI-Age.pdf

52 Trend, A. (2022). *AI think tanks to explore diversity and inclusion in Australian AI: Prof Didar Zowghi.* CSIRO Data61. https://algorithm.data61.csiro.au/ai-think-tanks-to-explore-diversity-and-inclusion-in-australian-ai-prof-didar-zowghi/

53 Zowghi, D. & Bano, M. (2022). *Diversity and Inclusion in Artificial Intelligence: Why?* LinkedIn Pulse. https://www.linkedin.com/pulse/diversity-inclusion-artificial-intelligence-why-didar-zowghi

54 A note on the lena image. (2018). Nature Nanotechnology, 13(12), 1087–1087. https://doi.org/10.1038/s41565-018-0337-2

55 Microsoft AI Team. (2022). *Seeing AI*. Microsoft. https://www.microsoft.com/en-us/ai/seeing-ai

56 Kelley, S. (2021). *Seeing AI: Artificial Intelligence for Blind and Visually Impaired Users*. VisionAware. https://visionaware.org/everyday-living/helpful-products/using-apps/seeing-ai-app/

57 Seeing AI (n.d.). *Microsoft Seeing AI and Low Vision Review.* Perkins School for the Blind. https://www.perkins.org/resource/microsoft-seeing-ai-and-low-vision-review/

58 Partnership on AI (2023). *Partnership on AI.* Partnership on AI. https://partnershiponai.org/

59 Inclusive Design Research Centre (n.d.). *The Inclusive Design Guide.* OCAD University. https://guide.inclusivedesign.ca/

60 Chou, J., Ibars, R., & Murillo, O. (2018). *In Pursuit of Inclusive AI.* Inclusive Design. Microsoft. https://www.microsoft.com/design/assets/inclusive/InclusiveDesign_InclusiveAI.pdf

61 Deloitte AI Institute (2021). *Opening doors of opportunity: AI as a vehicle for diversity and inclusion*. Deloitte Development LLC. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/technology/us-ai-institute-opening-doors-final.pdf

62 Ayadi, A. E. (2021). *Human-centred AI to build a trustful customer experience in retail.* https://medium.com/@alaeddineayadi/human-centered-ai-to-build-a-trustful-customer-experience-in-retail-f550b1b008

63 Nish P. (2021) *Understanding bias in AI-enabled hiring.* Forbes. https://www.forbes.com/sites/forbeshumanresourcescouncil/2021/10/14/understanding-bias-in-ai-enabled-hiring/?sh=5038fe2b7b96

64 Buonocore T. (2019). *Man is to Doctor as Woman is to Nurse: the Gender Bias of Word Embeddings.* Towards Data Science. https://towardsdatascience.com/gender-bias-word-embeddings-76d9806a0e17

65 Heaven, W. D. (2020). *Predictive policing algorithms are racist. They need to be dismantled.* MIT Technology Review. https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

66 Shin T. (2020) *Real-life Examples of Discriminating Artificial Intelligence.* Towards Data Science. https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070

67 Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–30. https://doi.org/10.1145/3359301

68 Pierce, A. (2020). *Measuring Fairness.* People+AI Research. https://pair.withgoogle.com/explorables/measuring-fairness/

69 Madaio, M., Stark, L., Wallach, H., & Vaughan, J. W. (2020). *Microsoft AI Fairness Checklist*. Microsoft Research. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4t6dA

70 Centre for Data Science and Public Policy (2018). *Aequitas - Bias and Fairness Audit Toolkit.* University of Chicago. http://aequitas.dssg.io/

71 Ruf, B. (2021). *The Fairness Compass: A Groundbreaking Step Forward for Trustworthy AI*. AXA. https://www.axa.com/en/insights/the-fairness-compass-a-groundbreaking-step-forward-for-trustworthy-ai

72 FairLearn contributors (2022). *FairLearn User Guide*. FairLearn. https://fairlearn.org/main/user_guide/index.html

73 Weinberger, D. (2022). *Playing with AI Fairness*. What-If Tool. https://pair-code.github.io/what-if-tool/ai-fairness.html

74 Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

75 Hussain, A. (2014). *Python Data Stack*. <packt>hub. https://hub.packtpub.com/python-data-stack/

76 IBM Research Team. (2022). *AI Fairness 360* [Source Code]. Trusted-AI. https://github.com/Trusted-AI/AIF360

77 IBM Research Trusted AI (n.d.). *AI Fairness 360 Community.* IBM Research. https://aif360.mybluemix.net/community

78 FairLearn contributors (2021). *Improve fairness of AI systems.* FairLearn. https://aif360.mybluemix.net/community

79 Amazon (2022). *Amazon SageMaker Clarify*. Amazon AWS. https://aws.amazon.com/sagemaker/clarify/

80 Mehrnoosh Sameki et al.. (2022). Microsoft Learn. https://learn.microsoft.com/en-us/azure/machine-learning/how-to-responsible-ai-scorecard

81 TrailHead (2022). *Use Einstein Discovery to Detect and Prevent Bias in Models.* Salesforce Inc. https://trailhead.salesforce.com/content/learn/modules/ethical-model-development-in-einstein-discovery-quick-look/use-einstein-discovery-to-detect-and-prevent-bias-in-models

82 Google People + AI Research (n.d.). *What-If Tool.* https://pair-code.github.io/what-if-tool/

83 Lee, M. S., & Singh, J. (2021). The landscape and gaps in open source fairness toolkits. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445261

84 Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., & Jacquet, A. (2022). Responsible AI Pattern Catalogue: A Multivocal Literature Review. https://doi.org/10.48550/arXiv.2209.04963

85 Reid, A., & O'Callaghan, S. (2018). *Ignorance isn't bliss*. Gradient institute. https://medium.com/gradient-institute/ignorance-isnt-bliss-6d133ee00f51

86 Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018). Learning Adversarially Fair and Transferable Representations. *arXiv*. https://doi.org/10.48550/arXiv.1802.06309

87 Vincent, M., & ManyOthers (2019). *Fairness.* scikit-lego. https://scikit-lego.netlify.app/fairness.html

88 Lattimore, F., O'Callaghan, S., Paleologos, Z., Reid, A., Santow, E., Sergeant, H., & Thomsen, A. (2020). *Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias*. Australian Human Rights Commission. https://humanrights.gov.au/sites/default/files/document/publication/ahrc_technical_paper_algorithmic_bias_2020.pdf

89 SecurityScorecard (2021). *What is the CIA Triad? Definition, Importance, & Examples*. https://securityscorecard.com/blog/what-is-the-cia-triad

90 The Difference between Security and Privacy and Why It Matters to your Program, https://www.hiv.gov/blog/difference-between-security-and-privacy-and-why-it-matters-your-program

91 The European Parliament and of the Council of 27 April 2016 (2016). *Regulation (EU) 2016/679 (General Data Protection Regulation).* Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

92 Australian Government (2022). *Privacy Act 1988.* No. 119, registered 16 November 2022. https://www.legislation.gov.au/Details/C2022C00321

93 Attorney-General's Department (2021). *Online Privacy Bill*. Australian Government. https://consultations.ag.gov.au/rights-and-protections/online-privacy-bill-exposure-draft/

94 IBM Research Trusted AI. (n.d.). AI Privacy 360. IBM Research. https://aip360.mybluemix.net

95 Salinger Privacy (n.d.). *Privacy Resources.* Salinger Consulting. https://www.salingerprivacy.com.au/publications/

96 Office of the Australian Information Commissioner. (n.d.). *Research and training resources*. https://www.oaic.gov.au/engage-with-us/research-and-training-resources

97 Australian Government Attorney-General's Department. (2022). Privacy Act Review Report. https://www.ag.gov.au/rights-and-protections/publications/privacy-act-review-report

98 Johnston, A. (2021). Privacy law reform in Australia - the good, the bad and the ugly. Salinger Privacy. https://www.salingerprivacy.com.au/2021/12/03/privacy-act-reform-proposals/

99 Swinson, M., Gunning, P., Evans, B., & Lim, C. (2023). Privacy Act Review Report (Finally) Released. King & Wood Mallesons. https://www.kwm.com/au/en/insights/latest-thinking/privacy-act-review-report-finally-released.html

100 Wong, C., Tsoi, K., Giral, M., Bhathela, N., & Knezevich, C. (2023). Australia's Privacy Act Review - Key Issues for Consultation. Hilbert Smith Freehills. https://www.herbertsmithfreehills.com/insight/australia%E2%80%99s-privacy-act-review-%E2%80%93-key-issues-for-consultation

101 Fai, M., Hii, A., Harris, C. (2023). Privacy Act Review Report: Highlights and Hot Takes. Gilbert + Tobin. https://www.gtlaw.com.au/knowledge/privacy-act-review-report-highlights-hot-takes

102 IBM Research Trusted AI. (n.d.) *Adversarial Robustness 360.* IBM Research. https://art360.mybluemix.net

103 Lincoln Laboratory (2022). *Responsible AI Toolbox.* MIT Lincoln Library. https://www.ll.mit.edu/r-d/projects/responsible-ai-toolbox

104 Vincent, J. (2016). *Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day.* The Verge. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

105 Taylor, J. (2023) GhatGPT's alter ego, Dan: users jailbreak AI program to get around ethical safeguards, The Guardian https://www.theguardian.com/technology/2023/mar/08/chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards

106 Amazon SageMaker (n.d.). *Amazon SageMaker Model Monitor*. Amazon AWS. https://sagemaker.readthedocs.io/en/stable/amazon_sagemaker_model_monitoring.html

107 Chen, S., Carlini, N., & Wagner, D. (2019). Stateful Detection of Black-Box Adversarial Attacks. *arXiv*. https://doi.org/10.48550/arXiv.1907.05587

108 Qin, Z., Fan, Y., Zha, H., & Wu, B. (2021). Random Noise Defense Against Query-Based Black-Box Attacks. *arXiv*. https://doi.org/10.48550/arXiv.2104.11470

109 European Data Protection Supervisor (n.d.). *D: Data minimization*. European Union. https://edps.europa.eu/data-protection/data-protection/glossary/d_en

110 Microsoft Presidio (n.d.). *Presidio: Data Protection and Anonymization SDK*. Microsoft. https://microsoft.github.io/presidio/

111 Kopp, A. (2021). *Create privacy-preserving synthetic data for machine learning with SmartNoise*. Microsoft Open Source Blog. https://cloudblogs.microsoft.com/opensource/2021/02/18/create-privacy-preserving-synthetic-data-for-machine-learning-with-smartnoise/

112 Radebaugh, C., Erlingsson, U. (2019). *Introducing TensorFlow Privacy: Learning with Differential Privacy for Training Data*. Tensorflow Blog. https://blog.tensorflow.org/2019/03/introducing-tensorflow-privacy-learning.html

113 Microsoft Research team (n.d.). *Microsoft SEAL*. Microsoft. https://www.microsoft.com/en-us/research/project/microsoft-seal/

114 TensorFlow (n.d.). *TensorFlow Federated: Machine Learning on Decentralized Data*. TensorFlow. https://www.tensorflow.org/federated

115 Microsoft Azure team (n.d.). *Azure confidential computing*. Microsoft. https://azure.microsoft.com/en-us/solutions/confidential-compute/#overview

116 Knott, B., Venkataraman, S., Hannun, A., Sengupta, S., & Ibrahim, M. (2021). CrypTen: Secure Multi-Party Computation Meets Machine Learning. *arXiv*. https://doi.org/10.48550/arXiv.2109.00984

117 Aether Transparency Working Group (2022). *Aether Data Documentation Template*. Microsoft. https://www.microsoft.com/en-us/research/uploads/prod/2022/07/aether-datadoc-082522.pdf

118 OpenRefine contributors (2021). *OpenRefine*. OpenRefine. https://openrefine.org/

119 Amazon SageMaker (n.d.). *Amazon Sagemaker Data Wrangler*. Amazon. https://aws.amazon.com/sagemaker/data-wrangler/

120 IBM (n.d.) *IBM InfoSphere QualityStage*. IBM. https://www.ibm.com/au-en/products/infosphere-qualitystage

121 YData Labs Inc (2022). *pandas-profiling 3.5.0*. Python Software Foundation. https://pypi.org/project/pandas-profiling/#history

122 DataPrep.ai (2022). DataPrep. SFU Database Group. https://dataprep.ai/

123 Data-centric AI (DCAI). (2023). *Data-centric AI Resource Hub*. https://datacentricai.org/

124 Data-centric AI (DCAI) (2021). *Neurips data-centric AI workshop*. https://datacentricai.org/neurips21/

125 Intel (2021). *Scaling AI and data science - 10 smart ways to move from pilot to production*. VentureBeat. https://venturebeat.com/ai/scaling-ai-and-data-science-10-smart-ways-to-move-from-pilot-to-production/

126 Canuma, P. (2022). *MLOps: What It Is, Why It Matters, and How to Implement It.* Neptune MLOps Blog. https://neptune.ai/blog/mlops

127 Visengeriyeva, L., Kammer, A., Bär, I., Kniesz, A., & Plöd, M. (n.d.). *Machine Learning Operations (MLOps)*. INNOQ. https://ml-ops.org

128 Oladele, S. (2022). *A Comprehensive Guide on How to Monitor Your Models in Production.* Neptune MLOps Blog. https://neptune.ai/blog/how-to-monitor-your-models-in-production-guide

129 Samiullah, C. S. (2020). *Monitoring Machine Learning Models in Production*. ChristopherGS. https://christophergs.com/machine%20learning/2020/03/14/how-to-monitor-machine-learning-models/#risks

130 TensorFlow (n.d.). *TensorBoard: TensorFlow's visualization toolkit*. Tensorflow.https://www.tensorflow.org/tensorboard

131 SageMaker (n.d.). *Amazon SageMaker Model Monitor.* Amazon AWS. https://sagemaker.readthedocs.io/en/stable/amazon_sagemaker_model_monitoring.html

132 Czakon, J. (2022). *Best Tools to Do ML Model Monitoring*. Neptune MLOps Blog. https://neptune.ai/blog/ml-model-monitoring-best-tools

133 Prometheus Authors 2014-2022 (2022). *Prometheus*. Linux Foundation & Cloud Native Computing Foundation. https://prometheus.io/

134 Grafana Labs (2022). *Grafana*. Grafana Labs. https://grafana.com/

135 Moss, E., Watkins, E. A., Singh, R., Elish, M. C., & Metcalf, J. (2021). *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest*. Data & Society. https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

136 Moss, E., Watkins, E. A., Singh, R., Elish, M. C., & Metcalf, J. (2021). *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest*. Data & Society. https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

137 Ofqual (2020). *Exceptional arrangements for exam grading and assessment in 2020*. Gov.UK. https://www.gov.uk/government/consultations/exceptional-arrangements-for-exam-grading-and-assessment-in-2020

138 Data.gov.nz (2018). *Algorithm Assessment - Agency Submissions: June-July 2018*. Data Gov NZ, https://www.data.govt.nz/assets/Uploads/Algorithm-assessment-agency-submissions-June-July-2018.pdf

139 Future of Life Institute (FLI) (n.d.). *The Artificial Intelligence Act*. The AI Act. https://artificialintelligenceact.eu/

140 City of Helsinki AI Register (2020). *What is AI Register?* City of Helsinki. https://ai.hel.fi/en/ai-register/

141 City of Amsterdam Algorithm Register Beta (2020). *What is the Algorithm Register?* City of Amsterdam. https://algoritmeregister.amsterdam.nl/en/ai-register/

142 Google Cloud (n.d.). *Model Cards*. Google. https://modelcards.withgoogle.com/about

143 IBM Research (n.d.). *AI FactSheets 360*. IBM. https://aifs360.mybluemix.net/

144 Microsoft AI. (2022). *Responsible AI*. Microsoft. https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6#primaryR9

145 OpenAI (2023). *GPT-4 System Card*. OpenAI. https://cdn.openai.com/papers/gpt-4-system-card.pdf

146 Ridley, A., Anderson, J., Bell, J., Glocer, A., Burrell, A., Burleigh, M., Shidi, Q. A., & Johnson, B. (2022). *Aether*. University of Michigan. https://aetherdocumentation.readthedocs.io/en/latest/contributing/core-team.html

147 Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). *Datasheets for Datasets*. *arXiv*. https://doi.org/10.48550/arXiv.1803.09010

148 Atlassian (2023). *DACI: Decision documentation*. Atlassian. https://www.atlassian.com/software/confluence/templates/decision

149 Microsoft GitHub contributors (2022). *Code With Engineering Playbook: Design Decision Log*. Microsoft. https://microsoft.github.io/code-with-engineering-playbook/design/design-reviews/decision-log/

150 O'Sullivan, C. (2020). *Interpretable vs Explainable Machine Learning*. Towards Data Science - Medium. https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48

151 Amazon Web Services, Inc. (2023). *Interpretability versus explainability*. Amazon. https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html

152 Molnar, C. (2022) *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.https://christophm.github.io/interpretable-ml-book/

153 Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv*. https://doi.org/10.48550/arXiv.1811.11839

154 Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv*. https://doi.org/10.48550/arXiv.1909.09223

155 People + AI Research (n.d.) *Language Interpretability Tool (LIT)*. Google Research - Language. https://pair-code.github.io/lit/

156 Captum. (2022). *Captum: Model Interpretability for PyTorch*. Facebook Inc. https://captum.ai/

157 Lundberg, S. M., Lee, S. -I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30*. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

158 Hastie, T., Friedman, J., & Tisbshirani, R. (2017). Partial Dependence Plots. In *The elements of Statistical Learning: Data Mining, Inference, and prediction* (2nd ed., Ser. Springer Series in Statistics, pp. 369–370). essay, Springer.

159 Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

160 Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939778

161 Google People + AI Research (n.d.). *What-If Tool*. Google. https://pair-code.github.io/what-if-tool/

162 Microsoft development team (2021). *Error Analysis: Identify & Diagnose Errors to Build Responsibly*. Microsoft. https://erroranalysis.ai/

163 IBM Research Trusted AI. (n.d.) *AI Explainability 360*. IBM Research. https://aix360.mybluemix.net/

164 Hase, P. and Bansal, M. (2020) *Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behaviour?* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics

165 Karimi, A.-H., von Kügelgen, J., Schölkopf, B., & Valera, I. (2022). Towards causal algorithmic recourse. *XxAI - Beyond Explainable AI*, 139–166. https://doi.org/10.1007/978-3-031-04083-2_8

166 Mitchell, E. et al. (2023) DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature, https://arxiv.org/abs/2301.11305

167 Lyons, H., Velloso, E., & Miller, T. (2021). Conceptualising contestability: Perspectives on Contesting Algorithmic Decisions. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 7. https://doi.org/10.1145/3449180

168 Kaminski, M. E., & Urban, J. M. (2021). The Right to Contest AI. *Columbia Law Review*, *121*(7), 1957–2048.

169 European Union (2020). *General Data Protection Regulation (EU GDPR) 2016/679*. 22(3). https://gdpr.eu/article-22-automated-individual-decision-making/

170 Lyons, H., Velloso, E., & Miller, T. (2021). Conceptualising contestability. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 9-13. https://doi.org/10.1145/3449180

171 Alfrink, K., Keller, I., Doorn, N., & Kortuem, G. (n.d.). *Contestable AI by Design*. Delft University of Technology. https://contestable.ai/

172 Almada, M. (2019). Human intervention in automated decision-making. Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law. https://doi.org/10.1145/3322640.3326699

173 Australian Financial Complaints Authority (AFCA) (2018). *Internal dispute resolution tips*. AFCA. https://www.afca.org.au/make-a-complaint/complain/internal-dispute-resolution-tips

174 Gardner, A. (2022). Responsibility, recourse, and redress: A focus on the three R's of Ai Ethics. IEEE Technology and Society Magazine, 41(2), 84–89. https://doi.org/10.1109/mts.2022.3173342

175 Queen's Printer of Acts of Parliament. (2018). *Data Protection Act 2018*. Legislation.gov.uk. https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

176 Loi, M., & Spielkamp, M. (2021). *Towards accountability in the use of Artificial Intelligence for Public Administrations*. Algorithm Watch & University of Zurich. https://algorithmwatch.org/en/wp-content/uploads/2021/05/Accountability-in-the-use-of-AI-for-Public-Administrations-AlgorithmWatch-2021.pdf

177 Novelli, C., Taddeo, M., & Floridi, L. (2022). Accountability in artificial intelligence: What it is and how it works. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4180366

178 Gradient Institute (2022). *Training*. Gradient Institute Ltd. https://www.gradientinstitute.org/training/

179 UTS (2022). *Short Course - Ethical AI: from Principles to Practice*. UTS. https://open.uts.edu.au/uts-open/study-area/Technology/AI--ML/ethical-ai-from-principles-to-practice/

180 Lute, C., Stoyanovich, J., & Verhulst, S. (n.d.) *AI Ethics*. AI Ethics: Global Perspectives. http://aiethicscourse.org/

181 Fast.ai (2020). *Practical Data Ethics*. Data ethics. http://ethics.fast.au/

182 Shankar, V., & Cook A. (n.d.). *Intro to AI Ethics*. Kaggle. http://kaggle.com/learn/intro-to-ai-ethics

183 Training & Certification (2022). *Ethics in AI and Data Science (LFS112x)*. The Linux Foundation. https://training.linuxfoundation.org/training/ethics-in-ai-and-data-science-lfs112/

184 Directorate-General for Communications Networks, Content and Technology (n.d.) *Regulatory framework proposal on artificial intelligence*. European Commission. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

185 Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2021). Conformity assessments and post-market monitoring: A guide to the role of auditing in the proposed European AI Regulation. *Minds and Machines*, *32*(2), 241–268. https://doi.org/10.1007/s11023-021-09577-4

186 Santow, E. (2021). *Human Rights and Technology Final Report.* Australian Human Rights Commission. https://humanrights.gov.au/our-work/rights-and-freedoms/publications/human-rights-and-technology-final-report-2021

187 NIST (2022). *AI Risk Management Framework.* NIST. https://www.nist.gov/itl/ai-risk-management-framework

188 ISO (n.d.). *ISO/IEC 23894: Information technology — Artificial intelligence — Guidance on risk management.* ISO. https://www.iso.org/standard/77304.html

The National AI Centre is building Australia's responsible and inclusive AI future.

For further information
National AI Centre
1300 363 400
+61 3 9545 2176
naic@csiro.au
csiro.au/naic

For further information
Gradient Institute
info@gradientinstitute.org
gradientinstitute.org