Australian Human Rights Commission

Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias

Technical Paper

GRADIENT INSTITUTE

Consumer Policy Research Centre

CSIRO DATA 61

CHOICE

ISBN: 978-1-925917-27-7

Communications Unit
Australian Human Rights Commission
GPO Box 5218
SYDNEY NSW 2001
Telephone: (02) 9284 9600 TTY: 1800 620 241
Email: communications@humanrights.gov.au.

Design and layout: Herbert Smith Freehills
Cover image: iStock images
Internal photography: iStock images

Authors:
Finn Lattimore, Simon O'Callaghan, Zoe Paleologos, Alistair Reid, Edward Santow, Holli Sargeant and Andrew Thomsen.

Acknowledgements

- Herbert Smith Freehills for the design and publication of this Paper.
- Prof Rosalind Croucher, President, Australian Human Rights Commission.
- Tiberio Caetano, Darren Dick, Sophie Farthing, John Howell, Dan Pagendam, Lauren Perry, Linda Przhedetsky, Natasha Rose, Phoebe Saintilan, Bill Simpson-Young, Lauren Solomon, Liz Stevens, Julia Steward, Susan Thomas Rajan, Leon Wild.
- Simon Rice and Toby Walsh for their peer review and advice on this Paper.

Technical Paper Partners

Australian Human Rights Commission is Australia's National Human Rights Institution. It is an independent statutory organisation with responsibility for leading the promotion and protection of human rights in Australia.

Gradient Institute is an independent non-profit that researches, designs and develops ethical AI systems, and provides training in how to build accountability and transparency into machine learning systems. It provides technical leadership in evaluating, designing, implementing and measuring AI systems against ethical goals.

Consumer Policy Research Centre (CPRC) is an independent, non-profit, consumer think-tank. We work closely with policymakers, regulators, academia, industry and the community sector to develop, translate and promote evidence-based research to inform practice and policy change. Data and technology issues are a research focus for CPRC, including emerging risks and harms and opportunities to better use data to improve consumer wellbeing and welfare.

CHOICE: Set up by consumers for consumers, CHOICE is the consumer advocate that provides Australians with information and advice, free from commercial bias. CHOICE fights to hold industry and government accountable and achieve real change on the issues that matter most.

CSIRO's Data61 is the data and digital specialist arm of Australia's national science agency. We are solving Australia's greatest data-driven challenges through innovative science and technology. We partner with government, industry and academia to conduct mission-driven research for the economic, societal and environmental benefit of the country.

This Technical Paper is part of the Australian Human Rights Commission Technology Project which is supported by four major partners:

Australian Government
Department of Foreign Affairs and Trade

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

HERBERT SMITH FREEHILLS

LexisNexis®

Contents

Edward Santow

Human Rights Commissioner

Australian Human Rights Commission

Commissioner's foreword

Artificial intelligence (AI) promises better, smarter decision making.

Governments are starting to use AI to make decisions in welfare, policing and law enforcement, immigration, and many other areas. Meanwhile, the private sector is already using AI to make decisions about pricing and risk, to determine what sorts of people make the 'best' customers... In fact, the use cases for AI are limited only by our imagination.

However, using AI carries with it the risk of algorithmic bias. Unless we fully understand and address this risk, the promise of AI will be hollow.

Algorithmic bias is a kind of error associated with the use of AI in decision making, and often results in unfairness. Algorithmic bias can arise in many ways. Sometimes the problem is with the design of the AI-powered decision-making tool itself. Sometimes the problem lies with the data set that was used to train the AI tool, which could replicate or even make worse existing problems, including societal inequality.

Algorithmic bias can cause real harm. It can lead to a person being unfairly treated, or even suffering unlawful discrimination, on the basis of characteristics such as their race, age, sex or disability.

This project started by simulating a typical decision-making process. In this technical paper, we explore how algorithmic bias can 'creep in' to AI systems and, most importantly, how this problem can be addressed.

To ground our discussion, we chose a hypothetical scenario: an electricity retailer uses an AI-powered tool to decide how to offer its products to customers, and on what terms. The general principles and solutions for mitigating the problem, however, will be relevant far beyond this specific situation.

Because algorithmic bias can result in unlawful activity, there is a legal imperative to address this risk. However, good businesses go further than the bare minimum legal requirements, to ensure they always act ethically and do not jeopardise their good name.

Rigorous design, testing and monitoring can avoid algorithmic bias. This technical paper offers some guidance for companies to ensure that when they use AI, their decisions are fair, accurate and comply with human rights.

On behalf of the Australian Human Rights Commission, I pay tribute to our partner organisations in this project for the deep expertise they provided throughout this work: Gradient Institute, Consumer Policy Research Centre, CHOICE and CSIRO's Data61.

Edward Santow
Human Rights Commissioner
November 2020

1 Executive summary

This technical paper is a collaborative partnership between the Australian Human Rights Commission, Gradient Institute, Consumer Policy Research Centre, CHOICE and CSIRO's Data61. We explore how the problem of algorithmic bias can arise in decision making that uses artificial intelligence (AI). This problem can produce unfair, and potentially unlawful, decisions. We demonstrate how the risk of algorithmic bias can be identified, and steps that can be taken to address or mitigate this problem.

AI is increasingly used by government and businesses to make decisions that affect people's rights, including in the provision of goods and services, as well as other important decision making such as recruitment, social security and policing. Where algorithmic bias arises in these decision-making processes, it can lead to error. Especially in high-stakes decision making, errors can cause real harm. The harm can be particularly serious if a person is unfairly disadvantaged on the basis of their race, age, sex or other characteristics. In some circumstances, this can amount to unlawful discrimination and other forms of human rights violation.

This paper describes the outcomes of a simulation. We have simulated a typical decision-making process and identified five scenarios in which algorithmic bias may arise due to problems that may be attributed to the data set, the use of AI itself, societal inequality, or a combination of these sources. We investigate if algorithmic bias would be likely to arise in each scenario, the nature of any bias, and consider how it might be addressed. The scenarios are framed around a consumer's interactions with an essential service provider that most people will deal with at some point—an energy company.

We principally consider fairness by reference to the protected attributes in Australian anti-discrimination law, such as sex, race and age. We then use three fairness measures—selection parity, equal opportunity and precision parity—as potential indicators of algorithmic bias or discrimination. Next, we apply mitigation strategies to address any algorithmic bias, and consider the effect of those strategies by reference to any change in the fairness measures and the overall accuracy of the decision making itself.

Each of the five scenarios explored in this technical paper highlights a protected attribute. It shows how algorithmic bias may pose a risk of unlawful discrimination under federal, state and territory anti-discrimination and equal opportunity laws.

This paper deliberately adopts a broad definition of algorithmic bias. We observe that algorithmic bias can result in unfairness, which in some situations can amount to unlawful discrimination or other forms of illegality. Businesses

with strong ethical principles and a concern for their reputation will seek to act fairly, and so it will be important to avoid algorithmic bias regardless of whether this always amounts to unlawful behaviour.

This approach allows companies to be proactive in identifying the human rights risks in how they use AI, and then to ensure that they address these risks by acting lawfully *and* responsibly. Responsible use of AI starts before the AI system is used in a live scenario. It requires rigorous design, testing and monitoring to ensure it is not affected by algorithmic bias. We offer some guidance aimed at improving fairness standards in the operation of AI systems, thereby reducing the risk of unlawful discrimination.

2 Introduction

2.1 Purpose and scope

There is rapid growth in the use of **artificial intelligence (AI)**[1] in decision making. This is fuelled by the promise that AI can increase the efficiency, accuracy and cost-effectiveness of many forms of decision making.

However, there are also risks. This paper explores how unfairness, and potentially unlawful discrimination, can arise through the operation of **AI systems** used in decision making. The paper discusses how these problems can be identified, and some steps that can be taken to address or mitigate the problems.

We simulate a typical scenario in which an AI system is applied to assist in making **decisions**. We use a synthetic data set and test the decisions produced by the AI system. We analyse the results from technical, human rights, and consumer rights perspectives.

Human rights should be considered at all stages of the development and use of technology, including AI. There are several factors, particularly in the design phase, that can engage individual or collective human rights. Therefore, we should ensure a rigorous assessment of these elements to address the risk of individual or social harms.

In accordance with the United Nations Guiding Principles on Business and Human Rights, businesses should be proactive in identifying the human rights risks or impacts of their activities and relationships.[2] Businesses should ensure the lawful and responsible use of the tools they design and deploy. This paper seeks to offer some guidance aimed at improving fairness standards in the operation of AI systems.

2.2 Contributing partners

This paper is the product of a collaborative partnership between the Australian Human Rights Commission, Gradient Institute, Consumer Policy Research Centre, CHOICE and CSIRO's Data61.

This paper highlights the importance of multidisciplinary, multi-stakeholder collaboration and cooperation. The complexity of issues relating to the use of AI in certain decisions requires the engagement of teams that offer insights from academia, government, non-profits and the private sector.

Each of the partner organisations has contributed its own resources in undertaking the work for this project. Gradient Institute led the technical work in this project. The Consumer Policy Research Centre's Research Pathways Program contributed funding to support some of the project's technical work.

3 Background and context

3.1 AI systems

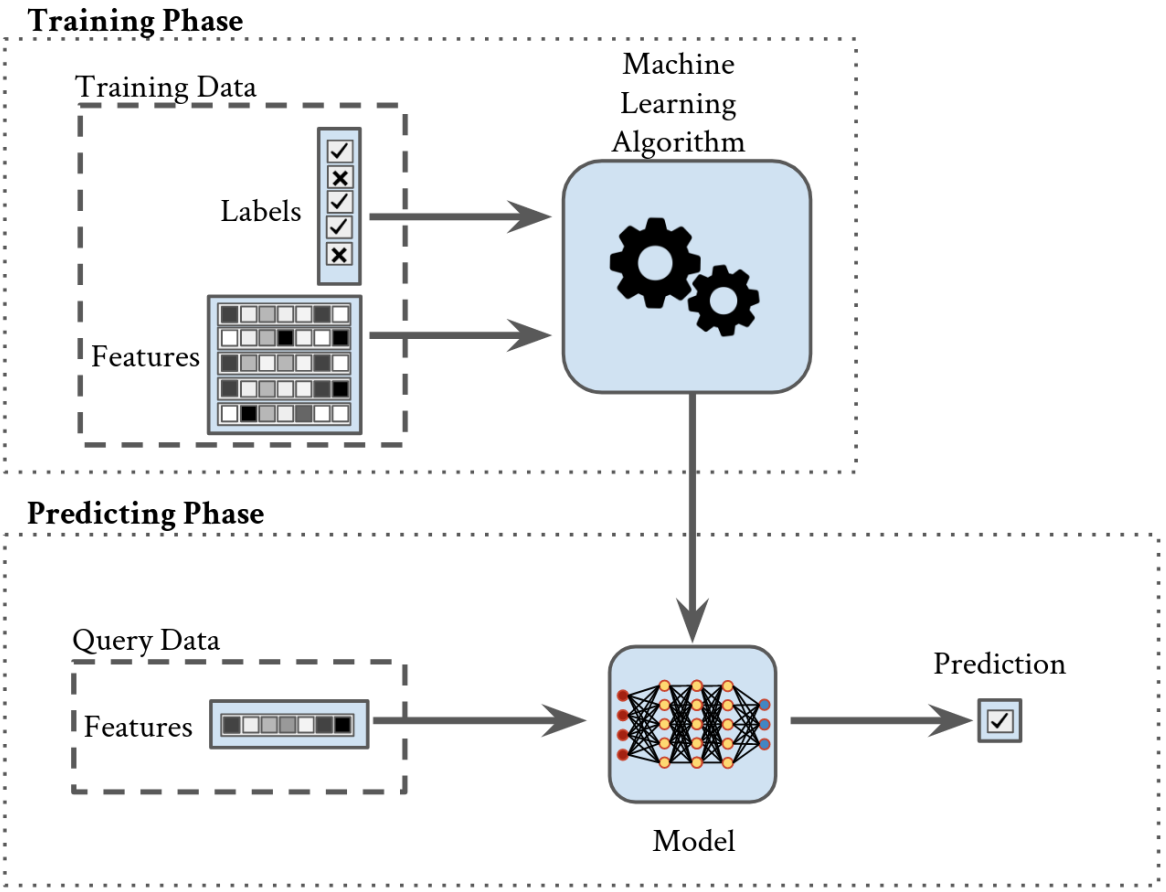
In this paper, the term 'AI system' refers to a system that materially uses AI in a decision-making process, whether or not the system is fully automated, or a person makes the final decision.

AI is not a singular piece of technology. Rather, it is a 'constellation of technologies'. There are generally considered to be two stages of development of AI: 'Narrow AI', which refers to AI systems capable of specific, relatively simple tasks; and 'Artificial General Intelligence' which is largely theoretical today and would involve sophisticated cognitive tasks. The AI system used in this paper employs Narrow AI that analyses data to develop solutions in a specific domain.

AI systems of the type discussed in this paper use data and **machine learning algorithms** to train mathematical **models** for the purposes of prediction or decision making, as shown at Figure 3.1. The **training data** commonly consists of previous decisions of the sort that the AI system will make or contribute to (**labels**) as well as supporting data (**features**).[3] The AI system then 'searches' for patterns within a data set of previous decisions, with a view to identifying common feature values or indicia associated with particular types of decision. Certain indicia might be present for past decisions that a company later assesses as 'good decisions' for the company's purposes, and other indicia might be present in decisions that the company subsequently judges to be 'bad'. Any future decision that relies on the AI system would be made by reference to a range of considerations that would include the identified patterns of previous decisions.

For example, an AI system that assists a bank in deciding whether to grant people home loans typically is trained on the bank's previous loan decisions, as well as any other data that the bank has access to. This can help the bank determine risk of default, by reference to an applicant's financial and employment history, and demographic information. In this way, the AI system can identify feature values or indicia associated with decisions to offer loans to people who turn out to be profitable (or unprofitable) for the bank. When the bank considers a new applicant for a bank loan (sometimes referred to as 'query data'), the AI system can be used to consider those feature values or indicia as they apply to the applicant, with a view to predicting whether the new applicant would be likely to pay back their loan reliably.

Figure 3.1: Overview of the type of AI system considered in this paper



This paper explores the risk that AI systems produce results that cause unfair disadvantage, or even unlawful discrimination, by reference to a hypothetical, but realistic, decision-making scenario. The aim is to examine the operation of a 'typical' AI system, with a view to understanding the human rights risks, and how these risks might be addressed.

The paper's hypothetical scenario considers how an electricity company, which is an example of an essential service provider, selects customers and potential customers using AI.

Specifically, the scenario involves an electricity company using an AI system to make decisions about whether to offer individuals (prospective customers) market-competitive service contracts. The AI system is provided with feature information related to each individual who is being considered for a service contract. These include data that such a service provider could reasonably obtain directly or from a third-party data broker. Some of this data may be predictive of the profitability of a customer. Sensitive variables, such as the individual's sex, race or age, may be included in the data set.

Even if not present, such sensitive information may be inferred by the AI system because the data set contains **proxy** variables which correlate with

the sensitive variable. The AI system is then used to predict each individual's likely profitability if they were accepted as a customer and, on that basis, the company decides whether to offer those individuals market-competitive service contracts.

3.2 Human rights framework

We consider the use of AI systems by reference to international and Australian human rights law.[4] AI systems can be used in ways that engage a number of human rights, but this paper focuses on one key right—namely, the right to equality and non-discrimination. This right is expressed in major international human rights treaties,[5] and it has been largely incorporated in Australian anti-discrimination laws.[6]

Some key human rights concepts for this paper include the following:

- **Equality** is predicated on the idea that all human beings are born free and equal. It means that all persons are equal before the law and are entitled without any discrimination to the equal protection of the law.[7]
- **Formal equality** is concerned with equality of treatment and expects all people to be treated the same way regardless of their differences.
- **Substantive equality** is concerned with equality of opportunity and outcomes. It recognises that formal equality does not address underlying, historical and structural inequalities that limit a person's opportunity to participate equally in society. Substantive equality goes beyond equal treatment, and attempts to redress underlying, historical and structural inequalities, which can require the use of affirmative action or 'special measures'.

Under international law, discrimination occurs when a person, or a **group** of people, is treated less favourably than another person or group because of their race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.

Discrimination can be direct or indirect. 'Direct discrimination' is where a person is treated differently from others. Discrimination also occurs when an unreasonable rule or policy applies to

everyone but has the effect of disadvantaging some people because of a personal characteristic they share. This is known as 'indirect discrimination'.

There are a series of laws that make discrimination unlawful at the federal level.[8] Those anti-discrimination statutes prohibit discrimination on the basis of **protected attributes**, including an individual's:

- age
- disability
- race, including colour, national or ethnic origin or immigrant status
- sex, pregnancy, marital or relationship status, family responsibilities or breastfeeding
- sexual orientation, gender identity or intersex status.

It can be unlawful to discriminate against a person on the basis of a protected attribute, in providing or refusing to provide goods, services or facilities.[9] Some state and territory anti-discrimination laws protect other attributes in their respective jurisdictions. Those other attributes include religion, immigration status, irrelevant criminal record, and profession, trade, occupation or calling.[10] While this paper focuses on anti-discrimination law, other laws can also be relevant to the operation of AI systems. For example, a range of federal, state and territory laws protect other human rights, such as privacy. In addition, Australian consumer protection frameworks aim to enhance the welfare of Australians by promoting fair trading, competition and consumer protections.[11]

3.3 Algorithmic bias in AI systems

(a) Algorithmic bias

Any decision-making system is capable of error. This is as true of decisions that are made using conventional methods that rely heavily on human involvement, as it is of the most highly sophisticated AI system, and every form of decision making in between. Where an AI system produces these sorts of errors, it is sometimes called algorithmic bias. Algorithmic bias is not a term of art; it is a general term that can refer to one or a collection of specific biases.[12]

This paper uses the term algorithmic bias to refer to predictions or outputs from an AI system, where those predictions or outputs exhibit erroneous or unjustified differential treatment between two groups.[13] The differential treatment may be erroneous because of mistakes or problems in the AI system, or it may be unjustified because it generally raises questions of unfairness, disadvantage or inequality that cannot be justified in the circumstances.

(b) Algorithmic bias, unfairness and unlawful discrimination

Erroneous or unjustified differential treatment can have particularly serious consequences if the groups are

distinguished by a 'protected attribute', such as disability, race, age or sex. This would include situations in which the outputs are different for people with a protected attribute in comparison with people without that protected attribute.

Algorithmic bias may be, or may lead to, unlawful discrimination, where there is no legal justification for that difference in **outcome**.[14] Addressing the problem of algorithmic bias in an AI system therefore will reduce the risk of engaging in unlawful discrimination. However, this technical paper deliberately adopts a definition of algorithmic bias that is *broader* than the strict legal definition of unlawful discrimination.

The reason for this approach stems from the fact that proving unlawful discrimination in a specific situation can be complicated, involving a detailed analysis of the particular facts and circumstances. Situations can arise that involve unfairness to a group, such as women or older people, where it is difficult or even impossible to prove in a court that they have suffered unlawful discrimination. However, very few companies would seek to act as unfairly as they could get away with, short of being sued for unlawful discrimination. On the contrary, any company with strong ethical principles and concern for its own reputation will seek to treat its customers and prospective customers fairly. Hence, for the vast majority of companies, it is important to avoid unfairness, regardless of whether their conduct ultimately could be proven also to involve unlawful discrimination.

(c) Sources of algorithmic bias

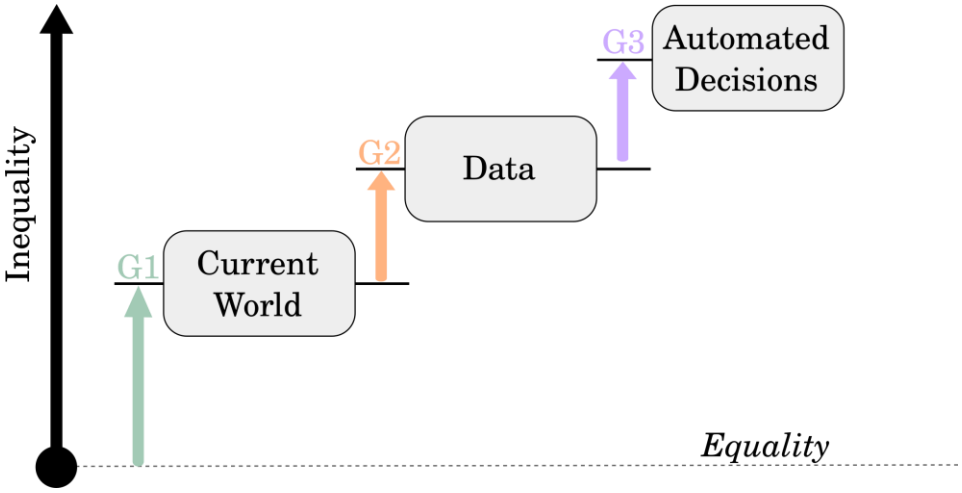
Another way of understanding the concept of algorithmic bias is to consider some of the root causes of inequality.

Figure 3.2 demonstrates inequality in the results of an AI system according to three components:

- G1 represents current societal inequality, extrinsic to the AI system
- G2 represents a source of algorithmic bias that may arise in an AI system due to inaccurate, insufficient, unrepresentative or outdated data
- G3 represents a source of algorithmic bias that may arise due to the intrinsic design or configuration of the AI system itself.

Algorithmic bias typically pertains to issues associated with G2 and G3, whereas societal bias with G1. The source of an AI system's inequality has implications for the effectiveness of any mitigation approaches taken.

Figure 3.2 Sources of inequality in AI systems.



For example, consider an AI system that is designed to predict individuals most likely to be profitable customers. Imagine the AI system disproportionately selects men, based on their predicted profitability, compared to women. There are various explanations for this unequal output, which fall broadly into two categories.

First, inequality may arise because of a problem with the data, or in the AI system design, or both. Unsuitable data (G2) or a poorly designed AI system (G3) can lead to predictions that are not representative of reality.

Depending on the precise detail of the situation, this algorithmic bias could also amount to unlawful discrimination, given that the differentiating factor was a protected attribute (whether the potential customer was male or female).

Second, inequality may arise where the AI system produces outputs that reflect existing inequalities external to the AI system. This may be due to societal inequalities such as the gender pay gap which has resulted in male customers actually being more profitable (G1). Although accurate, use of the AI system in this case would lead to a differential outcome for men and women.

It is important to note that gaps are not independent of each other. For instance, both G2 and G3 can reinforce or amplify G1 (an AI system that specifically disadvantages individuals who are already facing societal inequality).

3.4 Predictive modelling in consumer contexts

This paper demonstrates how algorithmic bias engages human rights in a consumer context through the provision of goods and services. Consumer rights are important to promote fair, safe and inclusive markets for Australian consumers, including equal access to products and services, which protects the human right to non-discrimination and equality.

We simulate a decision-making process that reflects common business practices where AI systems are used to predict the profitability (or creditworthiness) of a potential customer. AI systems and predictive modelling are often used to assist decision making in financial services, telecommunications, energy and human resources sectors.[15]

Businesses are increasingly collecting personal data, which they use in AI systems to improve the way they assess potential customers and ultimately increase their profitability. Often individuals have little choice about whether they are subjected to these almost-ubiquitous data-collection practices. If discrimination or other unfairness arises, this can breach their consumer[16] or human rights.

A company can pursue maximum profitability *provided they are not acting unlawfully*. In some situations, it would be unlawful to rely on an AI system that produces biased results. This is certainly true of an AI system that produces discriminatory results. Put simply, a business that makes decisions using an AI system that exhibits algorithmic bias faces a number of legal, financial and reputational risks that need to be carefully and conscientiously addressed.

3.5 Data sets

(a) Data sets and AI systems

AI systems are generally trained on large data sets. Some data sets contain personal information of many individuals. Other data sets are made up of personal information that has been aggregated or combined in a way that no individual's personal information is easily identifiable.

Where personal information is aggregated in a way that strips the detail linking it to an individual—often referred to as de-identified or anonymised data—the resulting data set will no longer be considered 'personal information' within the meaning of Australian privacy law.[17] Such aggregated data sets are frequently used in AI applications to draw inferences about individuals who share particular characteristics with groups of people—with that individual's personal information used as a reference point.

Digital platforms, such as Google and Facebook, collect personal data at large scale, and also offer AI systems that use predictive modelling. In addition, 'data brokers' buy, aggregate, analyse, package and sell personal data as well as insights derived from personal data.[18] This data can be used in AI systems.

(b) Data for purchase

Large data sets are commercially available for purchase. Australian and international companies offer these data sets for use in advertising, market research, insurance, financial services, health care, among other areas. These data sets consist of aggregated, anonymised user data gathered over a

period of time. These information-rich data sets provide access to credit risk scores, age, geography, debt balances, family circumstances, as well as political, social and consumer preferences and opinions, among other things.[19]

CHOICE contacted some companies to enquire about purchasing a data set.[20] CHOICE made some basic enquiries about their data set products and clients. These companies did not disclose how they collect and compile the information in these data sets.

The collection, aggregation and sale of large data sets present significant challenges for the businesses that rely on this information and use it in their operations. These practices may pose risks of harm to customers of those businesses.

(c) Simulated data

We use **simulated data** in this paper to study sources of algorithmic bias for two primary reasons. First, using simulated data, instead of the real personal data of a group of individuals, is the most effective way of protecting privacy. Secondly, real data sets will often contain some, if not all, the biases discussed in this paper. In order to adequately isolate and illustrate the effects of particular biases, we need the ability to *control* how these biases are introduced. This is only possible with a data simulator.

Details of the **simulation's** assumptions are discussed in each scenario as well as in Appendix 2.

4 Simulation

4.1 Retail electricity market and AI systems

Any business should ensure that its decision making is fair, accurate and avoids bias or discrimination. This proposition should be equally true for decision making that uses AI systems—hence the need to avoid algorithmic bias. While this paper assesses the risk of algorithmic bias in a scenario based on the retail electricity market, this is merely a hypothetical case study. The issues discussed in this paper apply to a variety of situations where AI systems are used to make decisions that have a significant effect on individuals, including insurance pricing, recruitment, mortgage lending and online marketing, to name just a few.

This simulation focuses on electricity service contracts because electricity is an essential service and some retail electricity providers assess prospective customers through credit checks.[21] These providers may plausibly use AI systems and data acquired directly or through a data broker to assess prospective customers.[22] This paper does not suggest that the simulation scenarios, including any particular algorithmic bias, unfairness or discrimination, describe the actions of any or all companies that participate in the Australian electricity market.

Australian electricity companies participating in a competitive market are generally not legally obliged to provide services to every prospective customer.[23] Companies routinely assess whether a potential customer is likely to be profitable, which involves predicting the cost of serving a potential customer alongside the revenue they might collect from the sale of services. This consideration can include the chance that the person will miss or make late payments, will need to access financial hardship provisions, will heavily use call centres and will require other support services. For example, a new entrant electricity company may have a business strategy built around offering low tariff prices to secure market share and minimise their own costs. To do so, they may aim to build a pool of customers who interact with the provider via cheap online services, pay reliably and require minimal additional support services, thereby excluding other consumers of whom many are likely to experience disadvantage. This will help reduce costs such as writing off debt, identifying and contacting consumers

who need support, administration of flexible payment options, and processes for upholding compliance with the regulations themselves.

If an electricity company assesses that a prospective customer is likely to encounter problems in paying their energy bills, the company may have little financial incentive to offer a market-competitive contract to that individual. However, receiving a cheaper market-competitive contract can also benefit an individual by making it easier for them to afford payments.

For the purpose of this simulation, we assume that selecting an individual for a market-competitive contract in this case is more beneficial to an individual, regardless of whether they will be profitable or not, because:

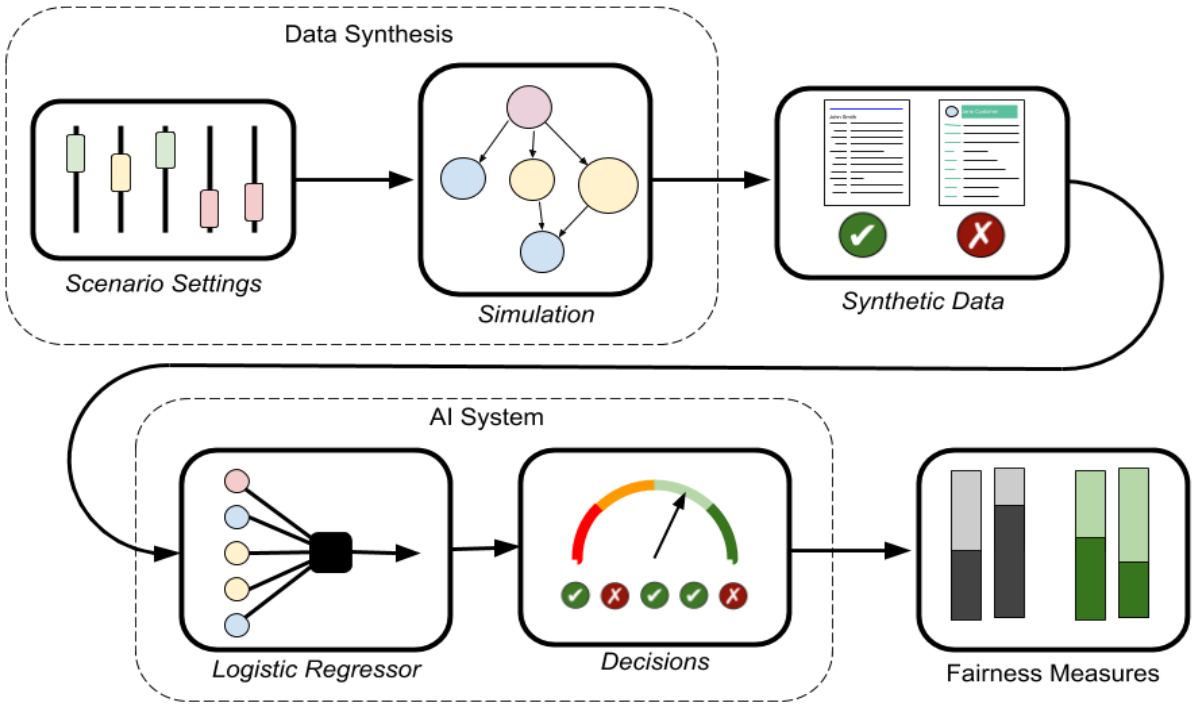
- electricity is an essential service and so individuals may settle for less advantageous deals or discounts, pay higher tariff prices, or receive poor customer service
- it would allow an individual to accept or reject the offer (though we acknowledge consumers may not be given adequate information to make an informed choice about whether an offer is in their best interests).

We are concerned about situations where individuals may not be offered a market-competitive contract due to bias or discrimination.

4.2 Data set synthesis and methodology

The process for the simulation conducted by Gradient Institute is outlined in Figure 4.1.

Figure 4.1 Illustration of the methodology used to investigate each scenario.



Scenarios and simulator

For each scenario, synthetic data is simulated containing customer records on which an AI system is trained to decide whether to offer a new applicant a market-competitive service contract. The data includes binary labels indicating whether each individual was profitable or not, which is used as the **target** for the AI system. If the AI system predicts a new individual to be profitable, a decision is made to offer the individual a service contract at market-competitive rates.

The simulated data set includes features that are known attributes for each individual associated with credit worthiness, including age, income, or postcode. Some of these features include protected attributes which demonstrate disparate outcomes between groups (such as sex, race or age). Additionally, this simulation includes **unmeasured features**, randomly generated, which may affect an individual's profitability, such as whether they pay bills on time, the amount of electricity they consume or if they use electricity in predominantly off-peak or on-peak times. The features and their dependencies on one another are, however, unique for each scenario and are set out in Appendix 2.

In each scenario, we use a sufficiently large data set to train the model, eliminating inaccuracies arising due to a small data set.[24] In simulating this data, we have made certain assumptions, such as the fraction of a population that is profitable or how predictive the data is, which would be difficult to measure in reality. In this simulation we are able to consider the question: *Would algorithmic bias be likely to arise in the outputs of an AI system trained on this data? If so, how might it be addressed?*

AI system

The synthetic data set is then used to train and validate a logistic regressor, which produces predictions of the chances that each customer will be profitable or not.[25] These predictions are then compared against a particular acceptance threshold to decide which customers are offered a market-competitive contract. Logistic regressors are widely used tools in the industry to address problems of **binary classification**. Binary classification is a modelling problem with the goal to distinguish between two categories. In this case, it is the answer to a yes or no question: *Is the individual going to be profitable?*

Logistic regression is a relatively simple mechanism but is sophisticated enough that our conclusions about resulting algorithmic biases generalise to a wider range of models.[26] Logistic regression makes a prediction by computing a weighted sum of the features and transforming the total through the logistic function. During the training of this AI system, the weighting of the features is adapted to optimise performance on the training data set so that the predictions match the **target variable's** values recorded in the training data (labels) as closely as possible.

Predictive modelling

In this case study, the AI system 'tunes' the weighting of each feature's contribution to the predictions, so that high likelihoods of profitability are allocated to individuals in the training data that are labelled profitable and low likelihoods of profitability are allocated to individuals labelled not profitable. Therefore, the AI system will be able to more accurately predict the target of new individuals for each scenario, assuming that the new individuals are statistically similar to the population from which the training data is obtained.

Feature importance

This weighting of the features is a measure of feature importance in a specific model which contributes to the AI system's prediction. However, the feature importance does not imply a causal relationship in reality. For example, if we attempt to predict whether an individual has the flu, knowing whether the individual performed a web search for flu remedies may be an important feature. However, it is not a causal feature. Preventing people from searching for flu remedies will not reduce the number of flu cases.

In the context of predicting profitability, a positive weight indicates that a high value of that feature is associated with profitable individuals, while a negative weight means the AI system will downrank those individuals as less profitable. Plotting these weights can help us understand the basis on which a model is making predictions.[27]

Fairness measures

We use three **fairness measures** to analyse algorithmic bias in relation to the outcomes of an AI system.[28] These fairness measures may indicate the potential presence of algorithmic bias or discrimination, but they are not determinative.

While these measures are framed in the context of 'parity', the intention is not necessarily to reach parity but to demonstrate the differences between two groups distinguished by a particular attribute.

- Selection parity (or demographic parity) compares the selection rates between groups. Selection parity requires the fraction of each group selected by the AI system (**selected individuals**) to be the same, regardless of differences in suitability for selection between groups. For example, an AI system that selects 40% of males and 30% of females for a job interview when an equal number of both genders apply, would fail selection parity.
- Equal opportunity compares the *correct* selection rate between groups, *considering only those who are **suitable for selection**.* If there is equality of opportunity, the chances of a suitable customer being selected will not depend on which group they belong to. For example, a scholarship program that selects 60% of suitable males and 80% of

suitable females might not constitute equal opportunity. (As a question of law, such a discrepancy still may be lawful if, for instance, the relative preferential treatment of female candidates could be shown to constitute a special measure that remedies historical inequality.)

- Precision parity compares the correct selection rate between groups, considering only those who are selected. A result complies with precision parity if the chances of a selected individual being suitable do not depend on which protected group they belong to. For example, a financial lending system in which 70% of the selected males were determined to be suitable and 50% of the selected females were determined to be suitable would fail precision parity.

4.3 Toolkit for mitigating algorithmic bias

Anyone who is considering the use of an AI system to make decisions should ensure that their decision making is fair and lawful. Government and businesses have particular obligations to their citizens and customers respectively. Where an AI system is used to make decisions that can affect a person's human rights, those decisions should be fair, lawful and they should uphold human rights.

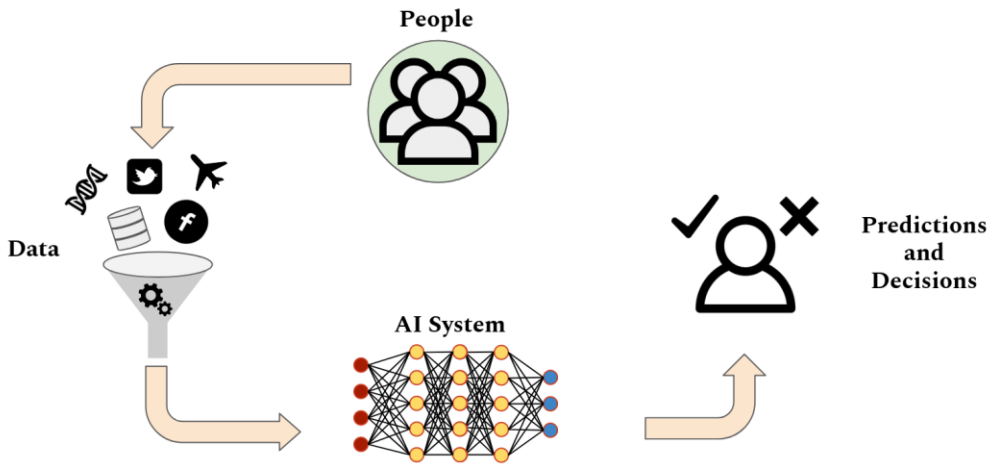
Fulfilling this responsibility starts before the AI system is used in a live scenario.

The AI system should be rigorously designed and tested to ensure it does not produce outputs that are affected by algorithmic bias. Once the AI system is operating, it should be closely monitored throughout its lifecycle to check that algorithmic bias does not arise in practice.

Using an AI system responsibly and ethically can extend beyond complying with the narrow letter of the law. Wherever an AI system causes harm, its use may not be responsible. For example, it is becoming clear that people of a lower socio-economic status often suffer disproportionately negative effects from algorithmic bias.[29] While Australian law does not expressly prohibit discrimination on the basis of socio-economic status,[30] avoiding this form of bias or less favourable treatment nevertheless would be good and ethical practice.

This simulation demonstrates that algorithmic bias may arise in AI systems in a consumer context, and that mitigation strategies may reduce the risk of these biases. An AI system can be designed and modified with positive steps to reduce algorithmic bias and societal inequality from being reflected in its outcomes. These steps, taken early in the AI design process to avoid algorithmic bias, can lead to more effective human rights protections.

Figure 4.2 AI systems in society

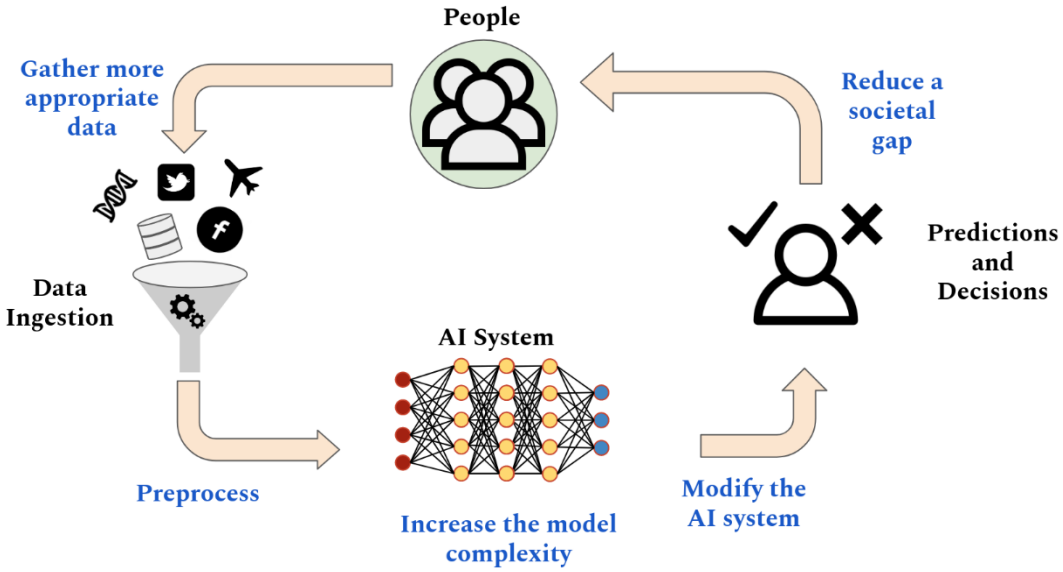


It is therefore important to continuously assess, test and mitigate algorithmic bias in an AI system. Mitigation strategies need to be tailored to the specifics of an AI system, appropriately address the potential sources of algorithmic bias, and carefully consider fairness metrics. There are several risk frameworks and strategies that businesses may adopt when considering these issues.[31]

This simulation highlights five general approaches to mitigating algorithmic bias. It is important for technical, policy, legal and other relevant decision makers to consider the specific circumstances of an AI system in its

context. These approaches are potential tools in a 'toolkit' of mitigation strategies—each strategy may be considered, tested and then only applied where it addresses the problem of algorithmic bias.

Figure 4.3 Mitigation strategies in AI systems



(a) Acquire more appropriate data

AI systems are increasingly based on data-driven modelling. It is important to understand the potential weaknesses and limitations of a data set that is required to train an AI system.

Data sets that are outdated, contain insufficient data points or insufficient characteristics or details about individuals can lead to inaccurate outcomes in an AI system. Individuals or groups that have faced systemic discrimination or are of under-represented groups may be further disproportionately affected by an AI system trained on poor data. When these problems arise, it is important to assess the risks of harm and reconsider the use of AI systems in the circumstances.

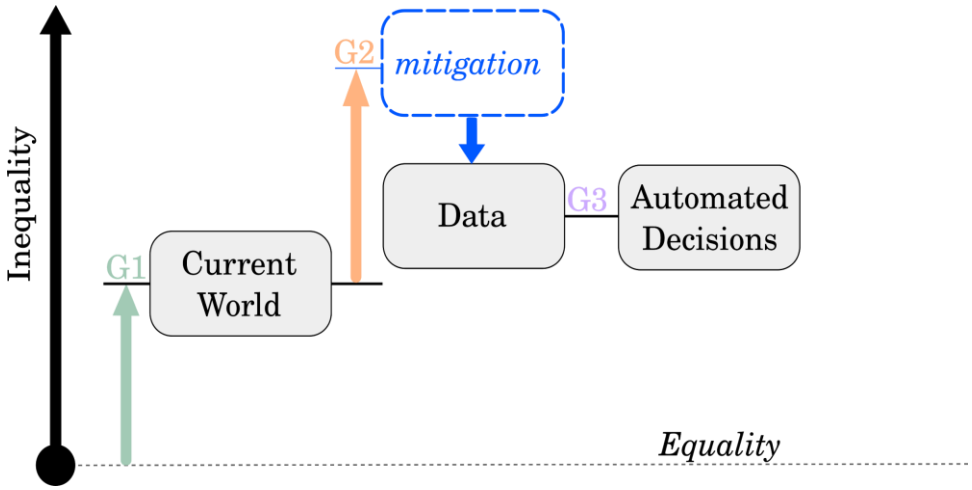
An important mitigation strategy is to responsibly obtain additional data points or new types of information relating to individuals inaccurately represented, or under-represented in the data set. This includes examples of historical bias in Scenario 2, label bias in Scenario 3, and under-representation in Scenario 4. Gathering additional or new data points reduces the G2 gap, as shown in Figure 4.4.

Even where an AI system may not result in discriminatory outcomes, historical bias may create inaccurate predictions by an AI system. Historical bias may arise over a number of decades of data or whenever the data is no longer

representative of reality due to a societal shift.

For example, the impacts of COVID-19 will be represented in demographic, financial and employment data collected in 2020. An AI system inclusive of this data impacted by the global health pandemic may have a number of implications, such as unfairly penalising people who had lost their jobs or suffered financial harm.

Figure 4.4 Acquiring more data about under-represented cohorts can help reduce the inequality between current and accurate data and the AI system's data set.



Considerations

This form of mitigation can be resource intensive, as it requires additional testing and trialling of outputs for groups at risk of experiencing algorithmic bias. It is often difficult to predict the benefits of additional data to the operation of an AI system before the new data is acquired and used.

The increased collection of personal information for data sets introduces additional human rights considerations. Collecting and using more information in a data set used to train an AI system may improve accuracy, but it may simultaneously limit the right to privacy of individuals whose personal information is collected and used in this way. Ultimately, increased data collection, including increased surveillance of all or parts of the population, can negatively affect the enjoyment of a number of human rights.

As discussed in Section 3.4, the additional risks and costs to a company

may be difficult to quantify but are important considerations. The increased cost associated with improving the data set by gathering more data or improving quality of information captured, may be greater than a company's potential increased profitability through use of an AI system. Additionally, companies may need to consider legal, financial or reputational risks. Deploying AI systems without rigorous testing on these under-represented or misrepresented **cohorts** may expose a company to legal risk of unlawful discrimination or failing to protect their consumers, as well as financial risk of incorrectly offering service contracts to consumers who are unable to meet their financial obligations.

Irrespective of these challenges, it is important that businesses take the steps necessary to ensure that individuals are not unfairly disadvantaged by AI systems. Additionally, taking these steps early in the AI lifecycle will ensure greater protections in the long term.

(b) Preprocess the data

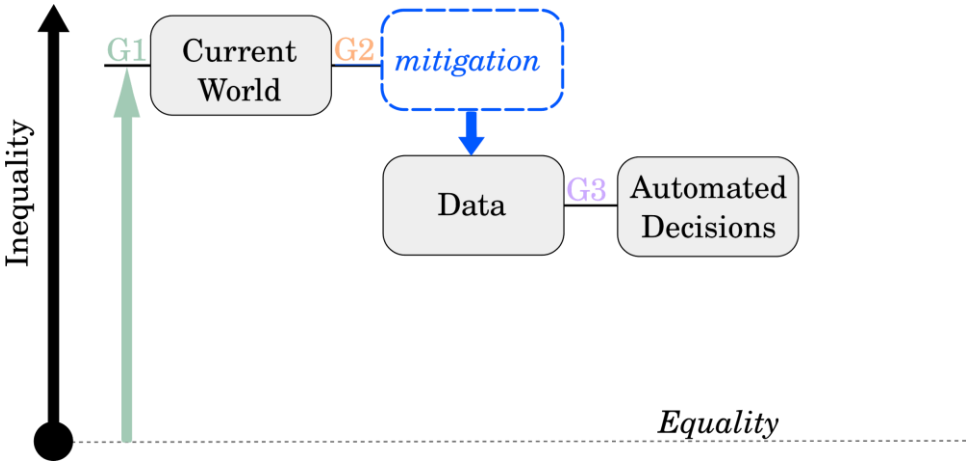
Preprocessing of data is a mitigation strategy that is used to edit features in the data set to mask or remove some information before it is used to train a data-driven algorithm.[32]

Preprocessing of data can hide a protected attribute like an individual's gender before it is used in the training data for an AI system to try to ensure no statistical difference in predictions between men and women. This may prevent individuals from being treated differently based on protected attributes, lowering the risk of algorithmic bias.

However, as with all mitigation strategies, it will be imperative to test the effect of hiding a protected attribute prior to deployment. In many cases, hiding protected attributes from the data set does not necessarily prevent these protected attributes from being considered indirectly by the AI system. Testing may identify that when a feature is deleted, other features such as employment, suburb, browsing and sales history act as a proxy variable for the deleted feature. An example of proxy variables encoding protected attributes is set out in Section 4.4 Scenario 2.

Figure 4.5 illustrates the potential impact of this mitigation on societal inequalities.

Figure 4.5 Preprocessing the data to partially mask protected attributes can reduce the effect of current world biases that target particular groups.



Considerations

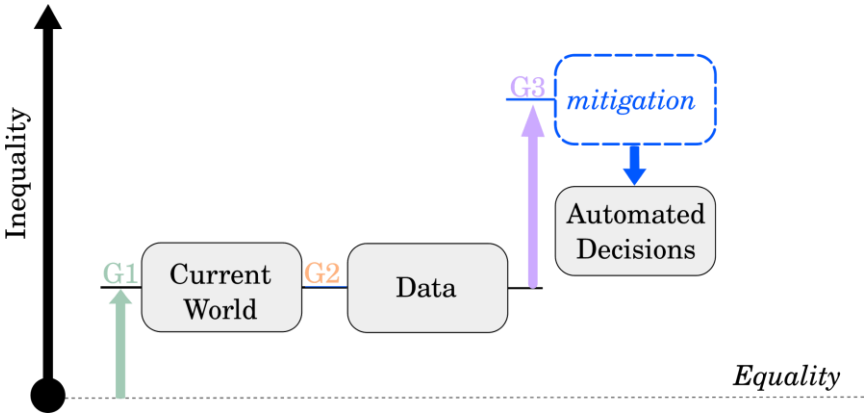
Preprocessing the data to hide protected attributes can also reduce an AI system's accuracy. This prediction error is more likely when the deleted protected attribute is closely correlated to the target the AI system is trying to predict. It is sometimes the case that mitigating inequality according to one measure of fairness may exacerbate the inequality measured by another (e.g. Scenario 1).

(c) Increase the model complexity

Simple models can be easier to test, monitor and interrogate. This makes them appealing to businesses seeking easier design and greater transparency of AI systems. Businesses should be aware of the potential that an over-simplified model will be less accurate. This can cause a model not to identify nuanced differences between groups (a G3 gap) which can lead to the model making generalisations that often favour the majority group.

However, the addition of new parameters increases the complexity of the model to identify and account for differences between groups in its predictions,[33] as demonstrated in Scenario 5. This will likely reduce potential algorithmic bias when appropriately trained, as compared with overly-simplistic models, while also increasing overall accuracy (illustrated in Figure 4.6). Testing the models on data sets prior to deployment will assist in identifying the impacts of complexity on accuracy and fairness.

Figure 4.6 Increasing the flexibility of a model can allow it to fit the data better and reduce inequality amongst populations where its inaccuracies were producing unfair outcomes.



Considerations

Increasing the complexity of the model comes at the cost of additional computing time and a potential reduction in model interpretability.[34] This may also increase the risk of overfitting the model, such that you reduce the overall accuracy of its predictions.

(d) Modify the AI system

An AI system may be designed or modified to correct for existing societal inequalities (a G1 gap), as well as other inaccuracies or issues in data sets causing algorithmic bias (a G2 gap). This is a simple approach to mitigating algorithmic bias, demonstrated in Scenario 1, but raises some important considerations.

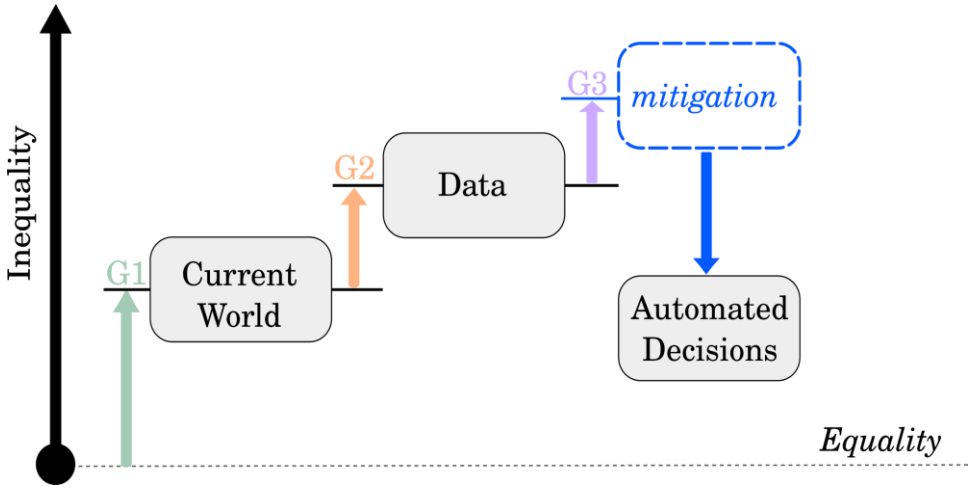
An AI system may consider a group that faces discrimination and, through testing, a company identifies that they are receiving lower scores or less preferable decisions, particularly if this reflects historical bias or existing social inequalities. To address this potential treatment, an AI system could be designed with a lower acceptance threshold for that group with the relevant protected attribute.

This approach adopts substantive equality—special measures may be taken to redress the inequalities between certain groups or individuals, as discussed in Section 3.2.

From a technical approach, implementing these modifications may involve:

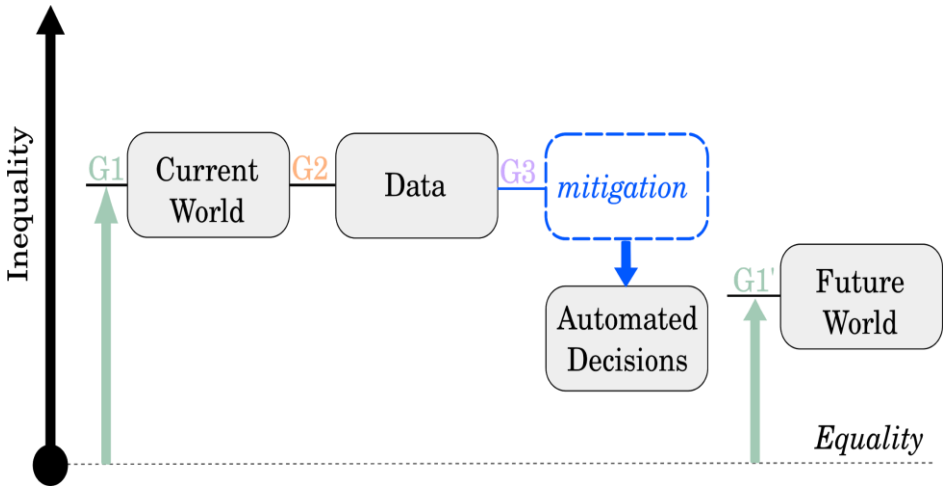
- Adjusting the decision logic applied to the prediction so as to favour a disadvantaged group.[35]
- Modifying the mathematical objective of the AI system so that, when learning from training data, it will choose a model that achieves a balance of fairness and accuracy.[36]

Figure 4.7 Mitigation through the modification of an AI system introduces a gap between the data and the automated decisions in an attempt to reduce inequality.



Even if the data does not introduce a new bias in the system, the AI system may be modified to remedy any differential treatment in predictions due to the existence of societal inequality (illustrated in Figure 4.8). Over time, these measures may contribute to improving outcomes for a disadvantaged group. The G1 gap is therefore reduced in the 'future world'.

Figure 4.8 Models with broad impact may be able to help reduce societal inequality over time by enacting decisions that are representative of a fairer society than the current one.



Considerations

Similar to preprocessing the data, improving one of the fairness measures with this mitigation strategy may affect another fairness measure (for example, see Section 4.3 Scenario 1).[37] Additionally, if the AI system is modified and designed to make decisions that do not reflect data collected from society (as demonstrated in Scenario 1), the accuracy of the AI system outputs may be reduced.

In this mitigation strategy, there may be some trade-off between improved outcomes for a disadvantaged group at the expense of an advantaged group.

(e) Change the Target

Questions like whether someone would make a profitable customer are complex because the concept of profitability has many elements. It is rarely possible to measure those specific elements quantitatively, let alone measure overarching concepts like profitability. As a result, we typically rely on a 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 that we are interested in may differ across groups. For example, using someone's credit history to predict their creditworthiness may be useful for older individuals with a track record of participating in the loan market, but it may disadvantage young people who are applying for their first loan. This difference between groups may result in unfair outcomes. Finding a fairer measure to use as the target variable would help alleviate this.

Considerations

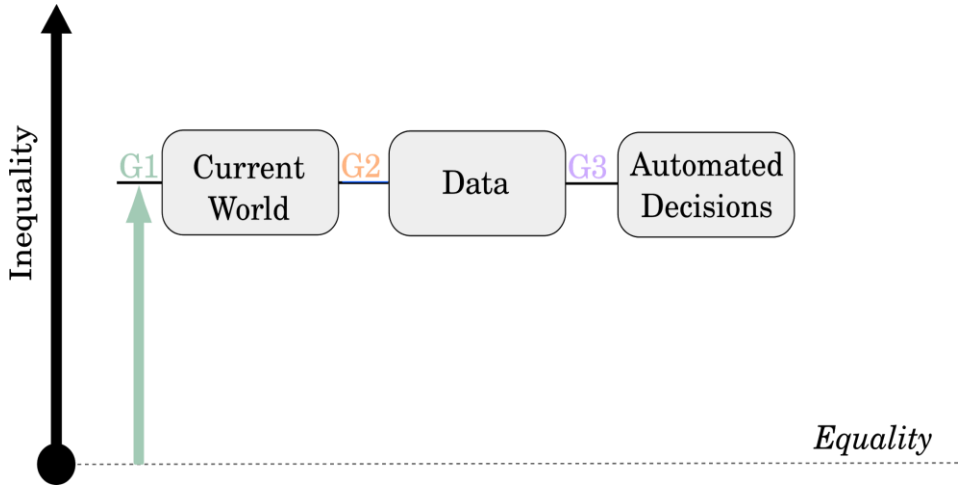
The lack of available data about the desired target, and the financial and practical constraints to obtain additional data sources, may make identifying another proxy variable infeasible.

4.4 Scenario 1: Different base rates

(a) Overview

Taken as a whole, Aboriginal and Torres Strait Islander peoples currently have lower average incomes than the broader Australian population.[38] Severe financial stress is present for half of Aboriginal and Torres Strait Islander peoples in Australia, compared with one in ten in the broader Australian population.[39] This scenario demonstrates the societal inequality between Aboriginal and Torres Strait Islander peoples and non-Aboriginal and Torres Strait Islander peoples, which may be reflected in an AI system (Figure 4.9 illustrates this inequality as a G1 gap).

Figure 4.9 Visual illustration of a G1 gap.



In this scenario, the simulated data accurately reflects the state of society and the AI system is capable of accurately modelling that data. Consequently, there are no issues in the data or in the design of the AI system (i.e. no G2 or G3 gaps), but the protected group is still disadvantaged.

If a protected group is predicted to be less profitable, due to lower incomes and endemic financial stress (a G1 gap), then even if we simulate or collect thorough data and train an accurate model in a way that does not introduce additional bias (no G2 or G3 gaps), the AI system will still perpetuate existing disadvantage by making decisions that mimic an imbalance in society. This can even happen indirectly when the protected attribute is not in the data set.

For this scenario, the model has simulated a data set that contains information about each individual's income, and the target being the profitability of the individual. We assume that an individual experiencing financial stress may be less profitable for the electricity provider, because the company needs to offer support such as deferred payments, repayment plans or increased support of a hardship program. This may increase a company's costs associated with hardship regulatory obligations, identifying and contacting consumers in need of support and administering these payment options. The AI system may use many additional features to predict a customer's profitability however we need only concern ourselves with income in this scenario to illustrate the potential for problematic system behaviour.

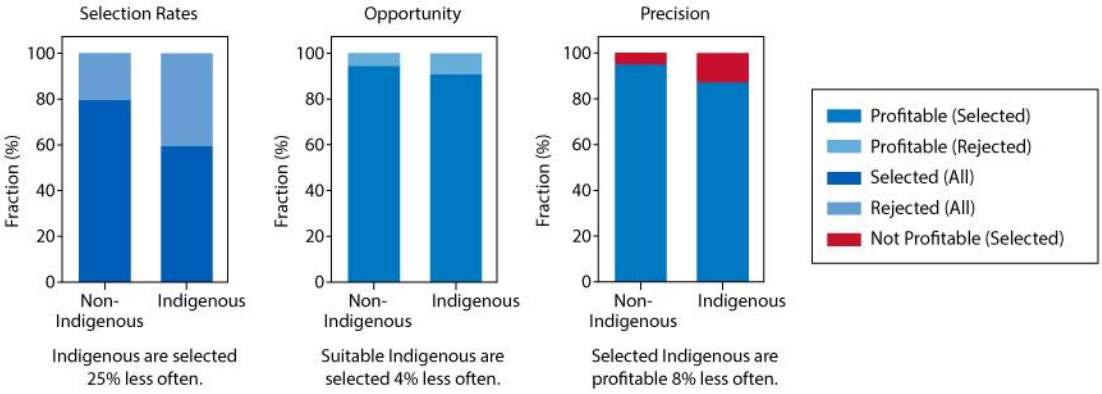
(b) Results

After training an AI system using this data set, we can examine its decisions with respect to its treatment of the protected groups.

One consequence of the historical and current disadvantage experienced by

Aboriginal and Torres Strait Islander peoples as a whole is a lower average income, compared with other people in Australia. Profitability is causally dependent on income, which therefore means a below-average income will likely result in lower profitability (see Appendix 2). The AI system could conceivably associate Aboriginal and Torres Strait Islander peoples with having below-average income, and therefore identify them as likely to be less profitable. This is regardless of whether a particular Aboriginal or Torres Strait Islander person happens to receive, in reality, an average or above-average income. (Figure 4.10, selection rates).

Figure 4.10 The selection, opportunity and precision rates for Aboriginal and Torres Strait Islander peoples and non-Aboriginal and Torres Strait Islander peoples.



In this scenario, the relevant feature that differentiates the groups (i.e., income) is known to the AI system. However, often there may be differences between groups that are correlated with the target but are not features that are inputted to the AI system. If this occurs, the AI system might use an otherwise redundant feature in the data set that correlates with a group as a means of improving predictive accuracy. For example, in this scenario, if income was not provided to the AI system, the model may assign high importance to postcode instead, which can correlate with regions where there is a high concentration of people with a low socio-economic status or of a particular ethnic or racial background. These substitutes are sometimes referred to as 'proxy variables' in the AI literature.

We will further discuss the concept of unfairness through redundant feature encoding in Scenario 2.

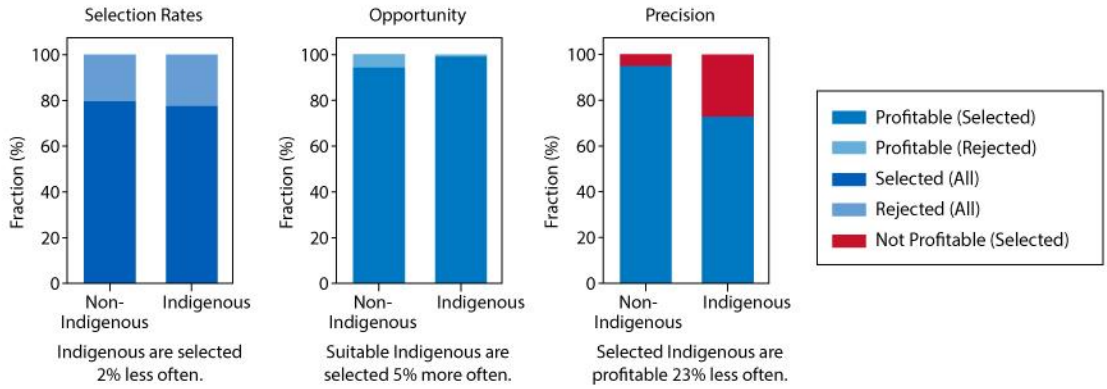
(c) Mitigation

A G1 gap in a specific AI system can be mitigated by interventions at the level of system design. Lowering acceptance thresholds for Aboriginal and Torres Strait Islander peoples compared to the rest of the population can be used to counterbalance the underlying differences in incomes and address the

selection parity. This form of mitigation is discussed in Section 4.3(d). Referring above to Figure 4.9, this would be equivalent to adjusting the model predictions in the AI system (a G3 gap pointing downwards) in order to address the societal inequality (a G1 gap).

An AI system that is effectively addressing societal inequality may over time improve the outcomes for a historically disadvantaged group. Widespread and sustained broader societal interventions are needed over time to ensure the societal gap is reduced or closed.

Figure 4.11 The fairness metrics after mitigation through post-processing of the model predictions.



Addressing one fairness metric can exacerbate another. In this case, lowering the acceptance threshold for Aboriginal and Torres Strait Islander peoples (levelling the selection rates) will result in an increase in accepted Aboriginal and Torres Strait Islander peoples who were not considered profitable as shown in Figure 4.11. Their debt obligations would likely need to be cancelled, and the costs subsumed by the company to mitigate this type of harm.

(d) Analysis of algorithmic bias and potential unlawful discrimination

This scenario is an example of a G1 gap which reflects existing societal inequality. The AI system is accurately selecting the more profitable cohorts, but in doing this it captures the broader structural inequalities connected with income and financial stress that have real world impacts on individuals' profitability for the service provider. The problems are not due to the data or the way in which the AI system has been trained.

This scenario also raises the risk that such decisions could contravene the *Racial Discrimination Act 1975* (Cth) (Racial Discrimination Act) or corresponding state or territory law. It

is unlawful for a provider of goods or services to discriminate against a person on the basis of race in refusing to provide a person with goods or services, or in providing goods or services on less favourable terms and conditions.

Where the effect of the operation of an AI system is that Aboriginal and Torres Strait Islander peoples are denied a service contract, or where they face additional barriers in accessing a service contract, this could be discriminatory under the Racial Discrimination Act.

In practice, whether such decisions in fact contravene the Racial Discrimination Act would involve a detailed legal analysis that is beyond the scope of this paper. But, for present purposes, it suffices to say that this scenario presents a risk that the Racial Discrimination Act may be breached. For legal and reputational reasons, it is imperative to identify and address such risks.

4.5 Scenario 2: Historical bias

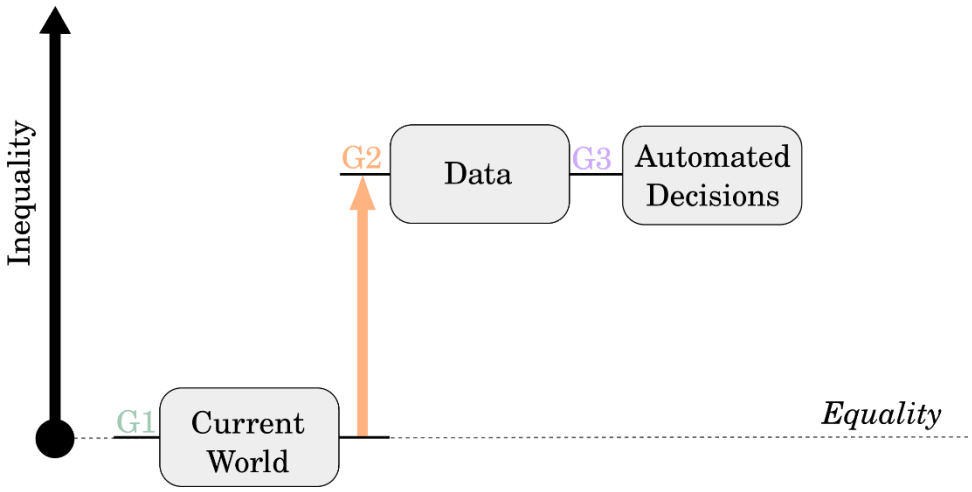
(a) Overview

Historical bias arises when the data used to train an AI system no longer accurately reflects reality. Women, as a whole, face a 'gender pay gap', barriers to leadership roles in the workplace and experience reduced employment opportunities due to family and caring responsibilities.[40] While women presently experience inequality in many areas of their lives, historically this inequality has been more pronounced without the significant progress of special measures supporting women in the workplace.[41]

To frame this scenario, we make two key observations. First, a person is not, as a question of objective fact, more or less profitable based on their sex. Second, historical inequalities between men and women have resulted in situations where women have appeared to be less profitable than men.

This scenario demonstrates the impact of training an AI system to decide whether or not an individual would be profitable based on historical data that is no longer reflective of the current world, introducing a G2 gap (Figure 4.12).

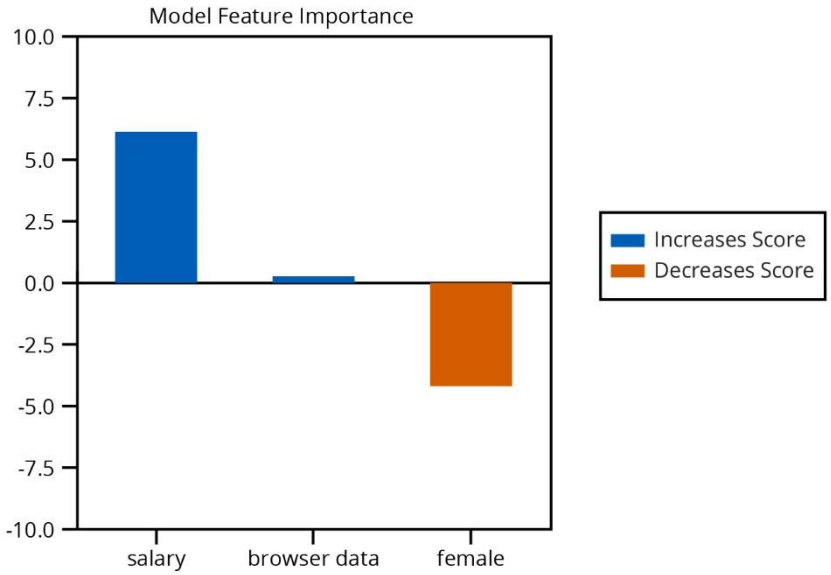
Figure 4.12 Historical data that is no longer representative of the current world introduces a G2 gap into the system.



(b) Results

Figure 4.13 shows that the model is placing a substantial negative weight on being female, which reflects historical discrepancies in the perceived profitability of women in comparison to men.

Figure 4.13 Feature importance for a model trained on historical data, where sex was predictive of profitability.

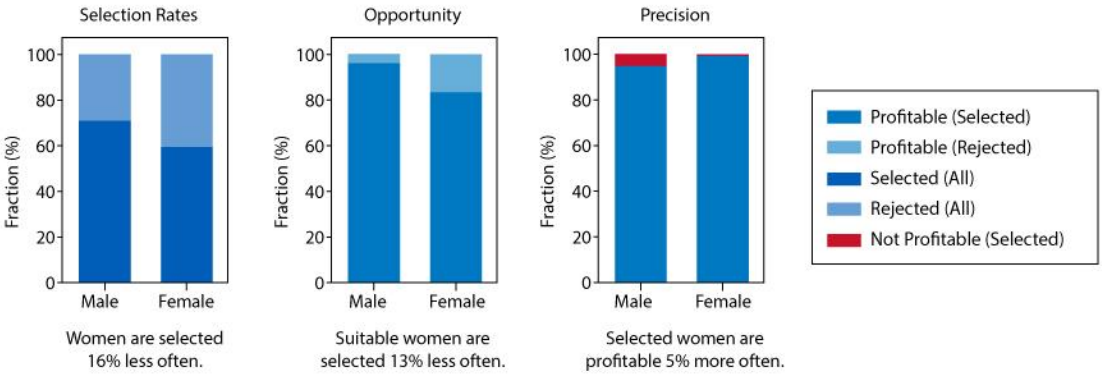


Additionally, the fairness measures in Figure 4.14 show that the AI system creates a substantial discrepancy with respect to selection parity and equality of opportunity. When comparing two groups—women who would be

profitable if selected and men who would be profitable if selected—13% fewer women are predicted by the AI system as profitable. This output is because the training data does not reflect the current reality that there is a reduced G1 gap between men's and women's income. The AI system is still analysing the world by reference to outdated historical data. Structural inequality outside the system (G1 gap) has reduced over time, but there is now a gap between the data and the current state of the world (G2 gap) because the AI system has been trained on information that is no longer current.

As the data does not reflect current reality, the AI system is denying service to women at rates that lead to discrepancies in selection parity and equal opportunity. The AI system also reduces profitability for the service provider as it does not select women who would in fact be profitable customers.

Figure 4.14 Fairness metrics for a model trained on historical data, where sex could be correlated to profitability.

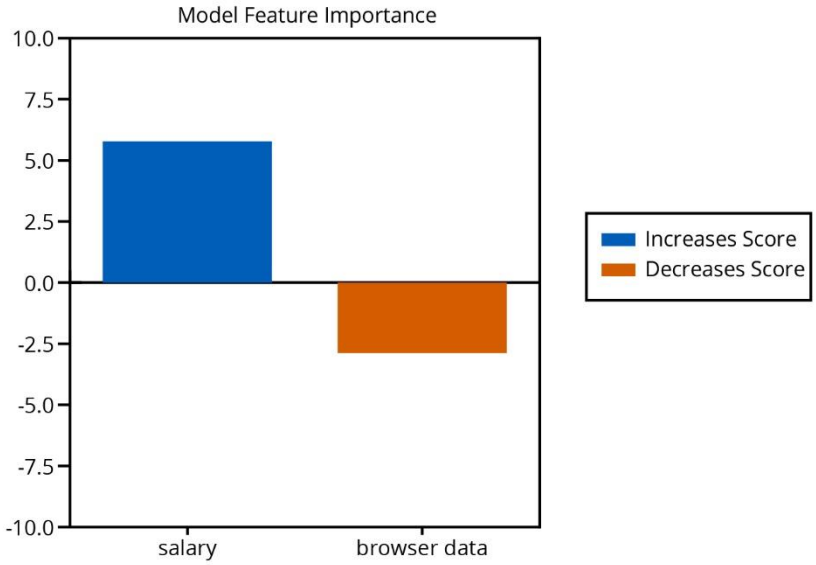


(c) Mitigation

To mitigate the discrepancy in the equality of opportunity fairness measure, one approach anecdotally employed by many businesses, is to remove the protected attributes from the data set with the aim of preventing an AI system from taking those protected attributes into account. However, this strategy may not work if the AI system takes account of a factor that acts as a proxy for the protected attribute. This will result in the AI system effectively still considering that protected feature.

Figure 4.15 shows the weight the model places on each feature in this scenario if sex is removed. The model now places a negative weighting for individuals with browser characteristics that are associated with being female.

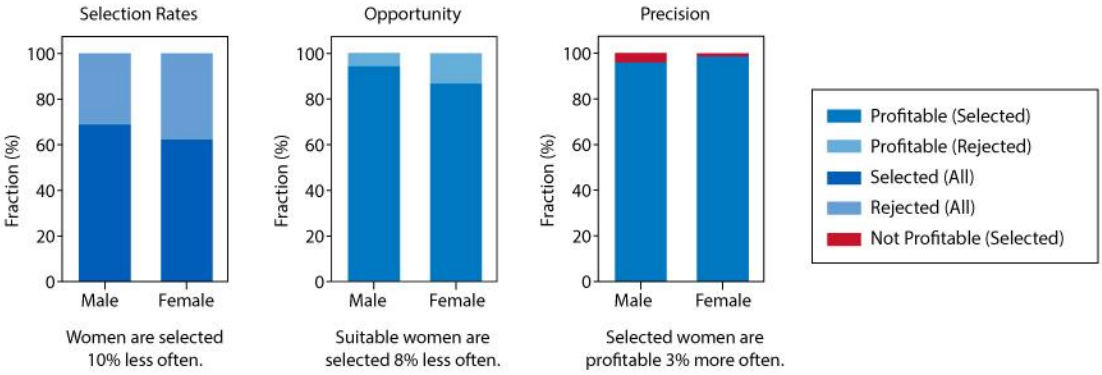
Figure 4.15 Feature importance weights if sex is removed from the data set.



Notice that browsing history, which was previously an insignificant feature, is now given a substantial negative weight. Browsing history can strongly correlate with sex,[42] so when sex is removed, the AI system uses browsing history as a proxy for sex, and downranks individuals who visit websites popular with women.[43] Sometimes a number of (otherwise benign) features can combine in such a way that they collectively become a proxy for a protected attribute. An AI system that considers the combination of otherwise benign features may have a similar effect as an AI system that directly considers a protected attribute.[44]

Figure 4.16 shows the fairness metrics with sex removed from the data set. The discrepancies with respect to selection rates and opportunity have reduced, because although browser history is acting as a proxy for sex, it does not perfectly correlate. The AI system is now penalising the group of people who visit websites popular with women, which includes some men. Similarly, women who predominantly browse websites associated with men will be included in the group predicted to be more profitable.

Figure 4.16 Fairness metrics after removing the sex column from the data.



Removing a protected attribute from the data set may reduce the algorithmic bias against a disadvantaged group. However, it may have only a limited effect if the data set is sufficiently information-rich (so as to include information that may act as a proxy) or can even increase the bias against a disadvantaged group (as discussed in Section 4.7).[45]

Ideally, impacts of historical bias would be mitigated by gathering a current data set that is representative of the current cohort. Figure 4.17 shows that if we retrain the model using a current data set, it no longer places any significant weight on sex (or a proxy, such as browsing data).

Figure 4.17 Feature importance after retraining the model with current data.

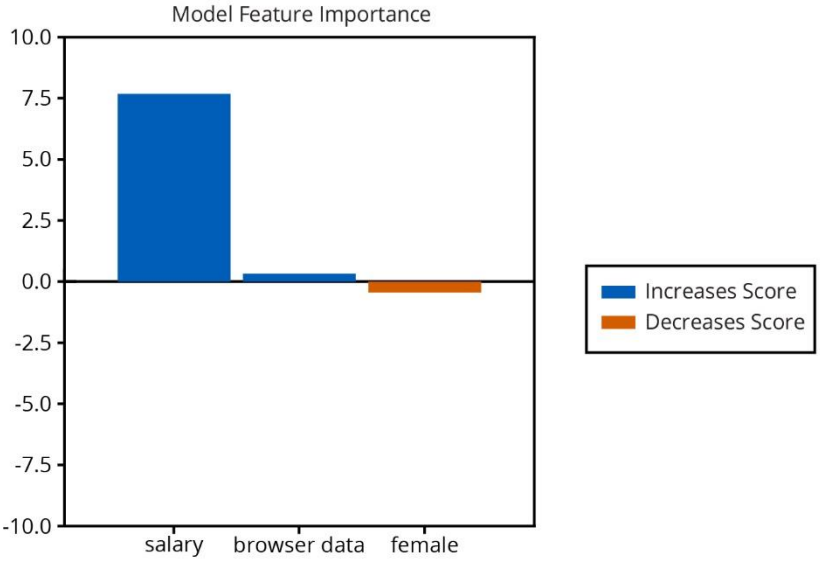
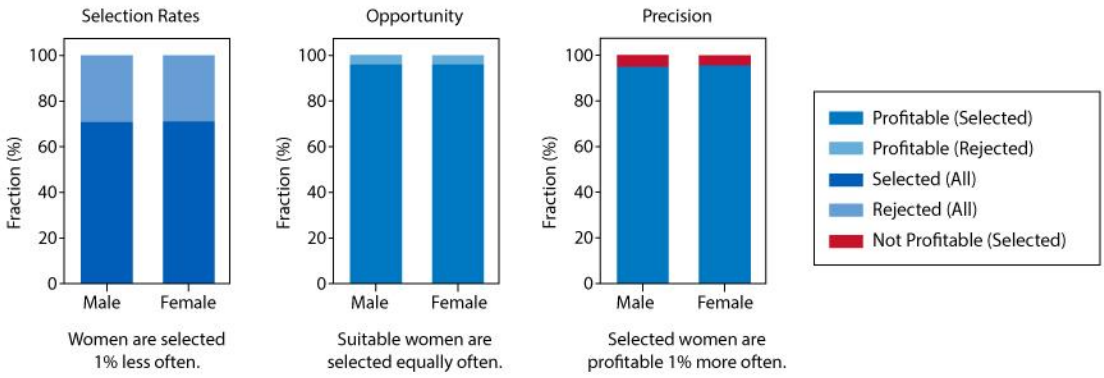


Figure 4.18 below shows that we obtain negligible discrepancies across the fairness metrics when the AI system uses more current data, rather than outdated historical data.

Figure 4.18 Fairness metrics after retraining the model with current data.



In practice, discarding historical data is often not practicable because it may render the available data set too small, decreasing the overall performance of the AI system.

As a technical alternative, it may be possible to model the change in the data over time, either theoretically or empirically, and the AI system's decisions could then be adjusted accordingly.[46]

(d) Analysis of algorithmic bias and potential unlawful discrimination

This scenario is an example of algorithmic bias where the predictions are not accurate due to the estimation of women's profitability based on historical bias, demonstrating a source of algorithmic bias from a G2 gap.

Where the effect of the operation of an AI system is to disadvantage women in obtaining a service contract, as described above, this could contravene the *Sex Discrimination Act 1984* (Cth) (Sex Discrimination Act), or corresponding state or territory law. Under the Sex Discrimination Act, it is unlawful to:

- treat women less favourably because of their sex or a characteristic that generally appertains to their sex
- impose a requirement on women, which is not reasonable, and which has the effect of disadvantaging some women on the basis of their sex
- discriminate on the basis of sex in refusing to provide a person with goods or services, or in providing goods or services on less favourable terms or otherwise unfairly.

This scenario presents a risk that the Sex Discrimination Act may be

breached. For legal and reputational reasons, it is imperative to identify and address such risks.

In taking steps to reduce this risk, it is worth noting that, within the machine learning literature, removing the protected attribute sometimes has been equated with avoiding direct discrimination.[47] This has encouraged designers to remove protected attributes when designing AI systems, because of the perceived importance of avoiding direct discrimination.

However, a more considered approach is needed, than simply removing reference to a protected attribute. It is not necessarily unlawful for an AI system to make decisions by reference to sex. For example, one relevant exemption for sex discrimination is where special measures are intentionally taken for the purpose of achieving substantive equality between men and women.[48]

Additionally, due to the ability of AI systems to draw inferences from a large number of subtle features, designers should be aware and informed about preventing indirect discrimination where proxy variables may still create less favourable outcomes for some people of a particular sex. This is particularly relevant considering the potential benefits of using a protected attribute in an AI system to mitigate against algorithmic bias.

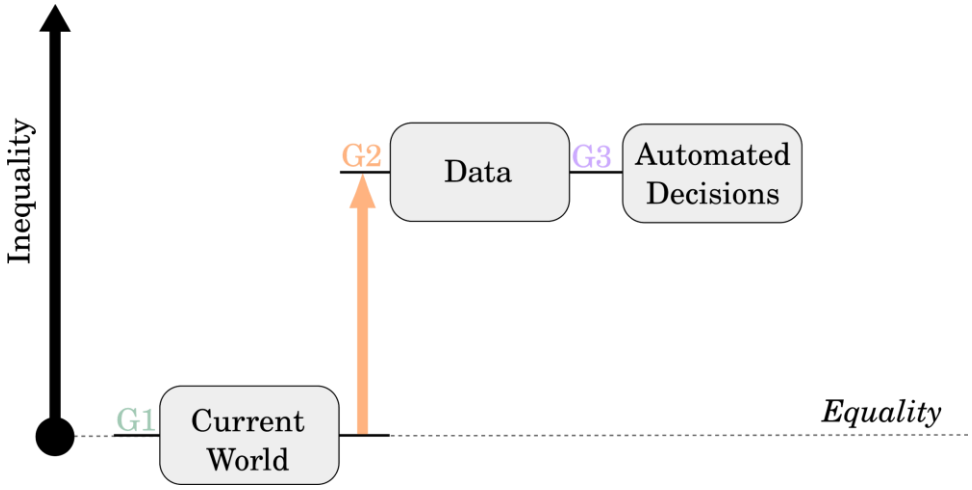
4.6 Scenario 3: Label bias

(a) Overview

Label bias may arise where there is a disparity between the quality of the label across groups that are distinguished by protected attributes (eg, age, disability, race, sex or gender). The label is the recorded value of the target that the AI system is trying to predict. In this scenario, the target is customer profitability.

For example, an AI system designed to predict profitability will be trained with labels that record whether previous customers were profitable or not. Label bias describes a systemic difference in the label accuracy of a particular group by virtue of a human bias in the recording of the target.

Figure 4.19 The biased labelling of the target introduces an inequality between the true state of the world and the data.



As shown earlier at Figure 3.1, the labels influence what mathematical model is produced by the system, which makes predictions. Preferential treatment for one group of individuals over another may artificially elevate or demote them in the data a company collects, which would then be reflected in the decisions made by any AI system trained on that data.

This simulation focuses on a scenario where one cohort is treated differently because of unconscious or conscious bias in customer service centre staff. Unconscious or conscious bias may manifest in a person's unfair treatment of another person, compared with the rest of the population.[49] Specifically, this hypothetical scenario demonstrates the potential for algorithmic bias where customer service centre staff treat customers from south-east Asian backgrounds less favourably than other customers.

As outlined at Section 4.1, these scenarios do not represent documented business practices in the electricity retail market but serve to illustrate potential pathways for algorithmic bias in decision-making processes.

South-east Asian Australians have been selected for this scenario because of a reported increase in racist treatment throughout the COVID-19 pandemic.[50]

One US study, conducted before the COVID-19 pandemic, showed that Caucasian customers generally received better quality service than black or Asian customers when requesting information from hotels through email communication.[51] That study was based on customer service support via written content (email contact) which included a personal name with strong race and gender associations. In the Australian context and for this scenario, a service centre staff member may associate an accent

or personal name with the individual having a south-east Asian background.

This hypothetical scenario illustrates how label bias could arise and is based on the following assumptions:

1. Any customer, regardless of their racial background, may experience financial hardship, and contact a customer service centre to request a special provision, such as an extension of time in which to pay their bill.

2. South-east Asian Australians are less likely to be granted special provisions, such as an extension of time to pay, as compared with other Australian customers, due to less favourable service and differential treatment based on their south-east Asian accent or name. To be clear, this is an assumption made for the purpose of this hypothetical scenario; this paper does not suggest that in reality electricity service providers disadvantage south-east Asian Australians in this way.

3. As a result of being less likely to receive special provisions when they experience hardship, south-east Asian Australians record higher rates of late payments and fees than other Australians, which is recorded in the data set.

4. Consequently, south-east Asian Australians appear to be less profitable according to customer data, even if in reality they are no more likely to miss bill payments.

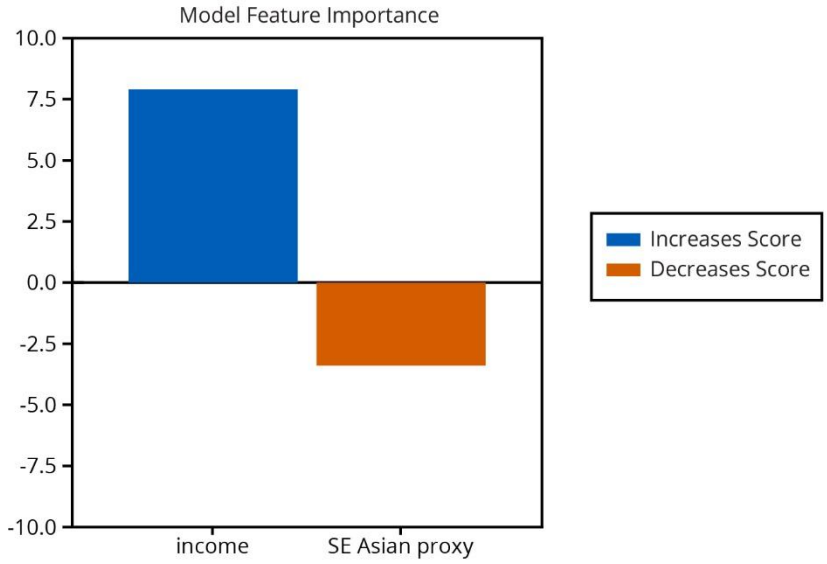
5. This label bias is correlated to south-east Asian Australian customers in the data set through proxy variable features which may be used to infer race (such as postcode or richer features available through third-party data brokers such as browser history). This is noted in Figure 4.20 as the 'SE Asian proxy'.

We consider the impacts of the AI system on south-east Asian Australians.

(b) Results

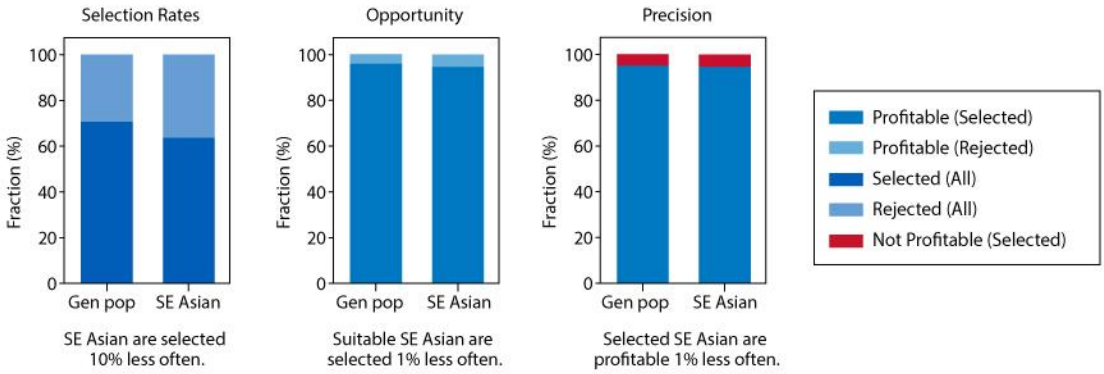
Figure 4.20 shows the effects that label bias in this scenario would have on the AI system. The AI system has mirrored the differential treatment based on race of the customer service employees by assigning a large negative weight to features that correlated with being Asian. Many features used by AI systems to predict the behaviour of individuals can correlate closely with race or ethnic origin, including postcode, occupation and browser history.[52]

Figure 4.20 Importance weights in the presence of non-SE Asian label bias.



Because south-east Asian Australian customers appear to accrue more late payments, the system will be less likely to identify as profitable future applicants from this demographic. The selection rate plot in Figure 4.21 illustrates that the biases of the employees result in the AI system selecting fewer south-east Asian Australians despite no "real world" differences (G1 gap) in their profitability compared to the rest of the population.

Figure 4.21 Fairness metrics in the presence of south-east Asian Australian label bias.



As the labels used to train and validate the system contain a human bias, detecting the distribution of error can be difficult. The AI system may appear to be equally accurate for both groups, given there is only 1% difference in equal opportunity and precision parity measures (shown in Figure 4.21).

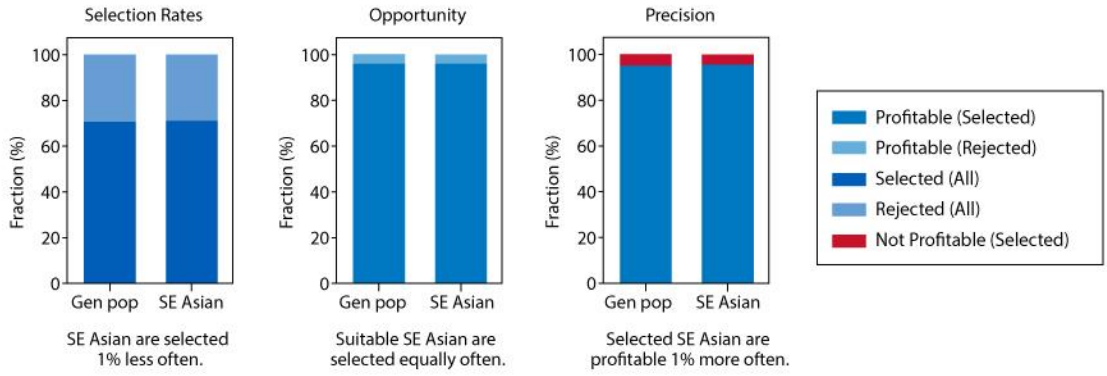
However, as the labels are affected by the prejudicial behaviour of the company's employees, the AI system has merely reproduced this view.

The selection parity fairness measure, which examines the proportions of different groups selected by the AI system, irrespective of how profitable they are, may give some indication of an underlying label bias issue (G2 gap) if a particular group is significantly over or under represented. However, this measure also includes the effects of different **base rates** between groups (G1 gap) as discussed in Scenario 1, so it may be difficult to identify the source of the bias.

(c) Mitigation

It may be possible to mitigate the problem of label bias by: reducing the prejudice affecting the label; using an alternative label that is not affected by the prejudice; estimating the prejudicial effect on the label and explicitly correcting for it; or, a combination of these steps.

Figure 4.22 Fairness measures when the label bias has been removed from the AI system.



In this case, label bias might be reduced by directly addressing the root cause of the prejudice. This could involve, for example, working to create cultural change within employees, so that they do not exhibit racial prejudice towards people who appear to be south-east Asian Australians.[53]

Alternatively, the degree of the prejudicial behaviour could be estimated by investigating the difference between the responses provided to south-east Asian compared with other Australians when they contacted support services. Sometimes it is possible to quantify the average level of label bias against a group, but not the bias associated with any individual label.[54] In these cases, the overall performance and equality of the AI system may both be improved by post-processing—that is, by lowering the decision threshold for south-east Asian Australians to compensate for their inflated risk scores.

(d) Analysis of algorithmic bias and potential unlawful discrimination

This scenario is an example of algorithmic bias where the predictions are not accurate due to the label bias in the data set towards south-east Asian Australians, demonstrating a source of algorithmic bias due to a G2 gap.

Due to this algorithmic bias and disproportionate disadvantage to south-east Asian Australians, there is a risk of unlawful discrimination. As discussed in relation to Scenario 1 at Section 4.3 (d), the Racial Discrimination Act, and corresponding state and territory laws, prohibit discrimination on the basis of a person's race, colour, descent, or ethnic or national origin.

There is a particular risk that an AI system affected by label bias could give rise to discrimination under the Racial Discrimination Act if it operates in a way that makes it harder for individuals of a certain racial or ethnic origin, such as Asian Australians, to access market-competitive service contracts or if the AI system imposes on them less favourable terms or conditions.

4.7 Scenario 4: Contextual features and under-representation

The predictions of AI systems are influenced by patterns and trends identified in the data across individuals. However, these patterns are not always transferable across groups and different demographics. For example, credit history may be a feature used by an electricity company's data-driven model to assess a new customer's ability to make their service payments. A lack of credit history may be associated with a customer who has faced financial problems, but it could also reflect circumstances where a customer is a young adult with no established credit history.

If the data does not capture the context for a customer's lack of credit history, an AI system may treat young adults the same way as it treats customers with poor credit history, and not offer young adults a market-competitive service contract.

Contextual features alone are not sufficient to cause algorithmic bias. If an AI system is sufficiently flexible, and has access to features that allow it (directly or indirectly) to identify young people, it can detect that a lack of credit history does not have bearing on profitability for young people and predict accordingly. However, if young people are either under-represented in the data (discussed in this Scenario 4), or the type of data collected does not adequately capture the behavioural difference between groups (see Scenario 5), discrepancies with respect to the fairness measures may arise.

(a) Overview

Algorithmic bias can arise towards a particular group of individuals with a protected attribute, where members of

that group are not adequately represented in the training data.

For example, an electricity company that has historically sold its products to an older demographic group may decide to expand its operations to target a younger audience. If they deploy an existing AI system that has been trained on their existing older customer data, the predictions of the AI system may systematically produce inaccurate predictions for the younger demographic group. This problem arises not from a real difference in the group's ability to pay bills, but because the feature values that correlate with a profitable older customer may not match with the feature values that correlate with a profitable younger customer. There is insufficient representation of younger individuals in the data for the AI system to learn the different behaviour between cohorts and consequently, the smaller group are subjected to a disproportionate erroneous decision.

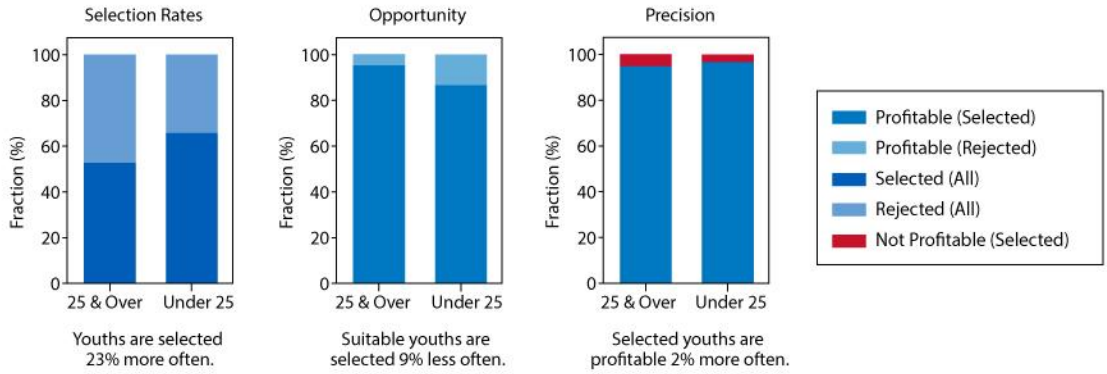
AI systems typically use training data to achieve a simple and general model. A general model that explains the majority of the data typically improves accuracy for predictions on new data. However, often this tendency to simplify comes at the expense of under-represented groups whom the AI system may ignore in its attempt to produce a model that performs well overall.

(b) Results

In this scenario, people under the age of 25 are significantly under-represented in the data set used to train the AI system, making up less than 1% of the 10,000 data points in the training set. As such, we would expect any correlations or patterns learnt by the AI system to produce less accurate predictions for the younger cohort.

Despite recording an accuracy of 94% on the training data, the accuracy drops considerably to 87% when used to decide which of the younger applicants to accept. The fairness metrics in Figure 4.23 also show that, despite a higher rate of under 25 year olds being accepted by the AI system, suitable young people are being accepted 9% less often. This is an indication of the AI system's decreased accuracy in predicting outcomes for this under-represented cohort.

Figure 4.23 Fairness metrics for a scenario in which the cohort of under 25 year olds is significantly under-represented in the training data.



(c) Mitigation

Gathering additional data from the under-represented cohort could help address this algorithmic bias. This would provide the AI system with more information about the behavioural patterns of the under-represented cohort while also increasing the penalty the system incurs if it ignores them. Acquiring an additional 900 data points (increasing the size of the under 25 year olds cohort to 10% of the total data set) improves the accuracy of the system on this group from 87% to 90%.

Good practice when designing an AI system is to first establish a baseline score for the method being used. Any decision to deploy the AI system should take into consideration how it compares to the baseline accuracy score. For example, if the AI system is less accurate than the current method, then its deployment needs to be justified against the benefits arising

from cost saving and its scalability for larger populations.

Additionally, the incentives of the AI system could be adjusted so it values all protected groups equally, regardless of the size of its representation in the data set. (The default configuration of typical machine learning algorithms,[55] places equal weight on every individual or instance of data with the result that good performance on small groups is not required to obtain good performance overall.) A drawback to this approach is that by placing a large weight on a small number of points, the influence of any statistical outliers in that group are amplified – potentially creating unpredictable behaviour in the AI system.

(d) Analysis of algorithmic bias and potential unlawful discrimination

This scenario is an example of algorithmic bias where the predictions are inaccurate due to the under-representation of data in relation to the under 25 year olds cohort. This scenario demonstrates the algorithmic bias that may arise where an inaccuracy does not necessarily create a selection rate disadvantage. In fact, in this scenario young people are selected 23% *more* often than the older cohort. However, suitable under 25 year olds (those who would be profitable if selected) are 9% less likely to be picked than a suitable applicant who is over 25.

Therefore, it is important to identify the potential risks of having inaccurate AI systems making decisions, particularly considering the potential disadvantage to groups who are likely to be under-represented in the data.

This scenario also raises a risk that such decisions could contravene the *Age Discrimination Act 2004* (Cth) (Age Discrimination Act), or corresponding state or territory law. Under the Age Discrimination Act, it is unlawful to:

- treat a person less favourably because of their age or a characteristic that generally appertains to their age
- impose a requirement on a person, which is not reasonable, and which has the effect of disadvantaging some people on the basis of their age
- discriminate on the basis of age in refusing to provide a person with goods or services, or in providing goods or services on less favourable terms or otherwise unfairly.

A detailed legal analysis would be necessary before any conclusions could be drawn about the AI system.

4.8 Scenario 5: Contextual features and inflexible models

(a) Overview

Issues may also arise in the presence of contextual features if the data contains insufficient information to capture the

differing behaviour of the various demographics (a G2 gap). Assuming a homogeneity in behaviour across groups can often result in a reduction in prediction accuracy—especially for under-represented groups—due to this averaging effect. We illustrate this by examining the same situation as Scenario 4, but limiting the available features being fed to the AI system by removing the protected attribute of age and all proxy features for age from the AI system. This prevents the AI system from attributing different behaviours in the data to the demographic groups to which they belong, regardless of the number of data points.

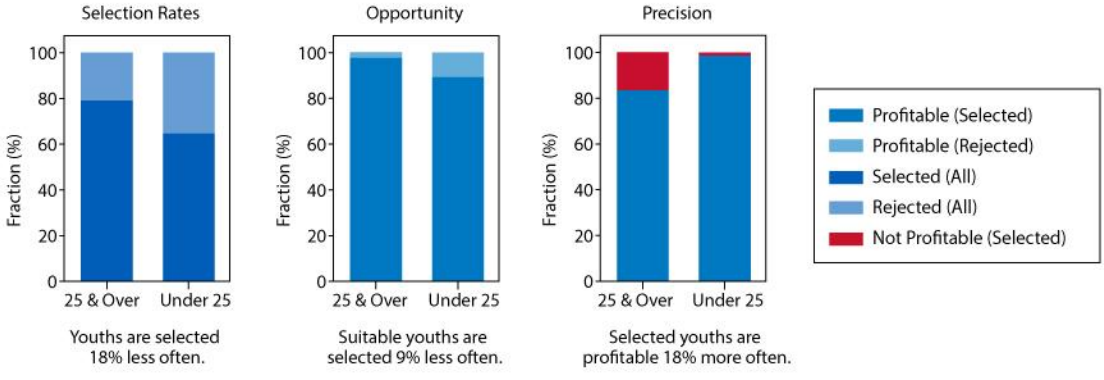
(b) Results

Figure 4.24 shows the fairness metrics when both groups have a more than sufficient amount of data points in the training set but the AI system is unable to tailor its predictions to each group's behaviour, due to a lack of information to otherwise be able to predict which group an individual may belong to.

Despite the prevalence of suitable customers being approximately similar for both groups, we see that the system gives preferential treatment to the over 25 year old group. This is likely due to the fact that, in this scenario, this group's suitability is closely linked to their income, whereas the suitability of under 25s depends more on unobserved factors. Unfairness is also apparent in the opportunity and precision metrics.

The equal opportunity measure shows that a disproportionate number of young people are erroneously not selected by the system. Meanwhile, the precision parity measure shows the AI system is accepting a disproportionate number of over 25 year olds who are not profitable, which may possibly result in the need for cross-subsidisation from the under 25 year old group.

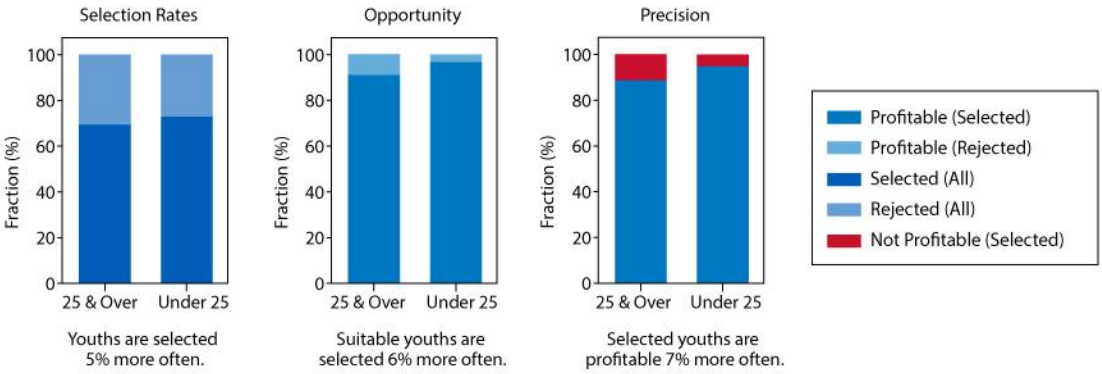
Figure 4.24 Fairness metrics where under-representation means the differing behaviour of separate demographics is not captured.



(c) Mitigation

Adding the sensitive attribute back into the AI system enables it to adjust each prediction depending on the group. Overall accuracy is increased from 88% to 90% and the fairness metrics are closer to parity (Figure 4.25). In some cases, where the groups' relationships between the features and the outcomes are sufficiently different, the AI system will perform better if a completely separate model is trained for each group.[56]

Figure 4.25 Fairness measures where a unique model is trained for each group to capture differing behaviour.



In this case, unlike the mitigation in Scenario 2, removing the protected attribute creates, rather than reduces, algorithmic bias. This highlights the fact that removing a protected attribute ('fairness through unawareness') is rarely an effective approach to achieve the desired intent of reducing unfairness.

(d) Analysis of algorithmic bias and potential unlawful discrimination

This scenario is a similar example of algorithmic bias where the predictions are inaccurate due to the under-representation of data in relation to the cohort of under 25 year olds. This scenario demonstrates algorithmic bias arising from a G3 gap.

It is important to consider the potential implications of this type of algorithmic bias in an AI system in relation to the Age Discrimination Act (see discussion in Section 1.1 (d). Algorithmic bias resulting from inaccurate predictions presents a risk of unlawful discrimination.

Additionally, it may be necessary to alter the design of an AI system to produce equal outcomes for under-represented groups. Designing different models for different individuals may increase accuracy and therefore improve fairness measures, particularly where the features have different implications for different

groups. However, if separate models are designed to predict the same outcomes for different individuals, there is likely to be a disparity between the outcome for an individual depending on which model is used.

The intention of this mitigation strategy is to take positive steps to create equal outcomes between two groups.

As discussed in Section 3.2, special measures may be taken to achieve substantive equality by redressing the inequalities between certain groups or individuals. In particular, the Age Discrimination Act provides an exemption to unlawful discrimination for measures intended to reduce a disadvantage experienced by people of a particular age.[57] Creating an AI system designed for individuals in an under-represented group, distinguished by a protected attribute, may fall within this exemption.

5 Charting a way forward

5.1 Risks of harm in predictive decision making

It is natural and legitimate that companies want to understand their current customers and prospective customers. This can enable companies to tailor product and service offerings appropriately, which ultimately can improve their profitability. How companies seek to develop this understanding is important. There are risks in developing an erroneous understanding of particular individuals—especially where errors unfairly disadvantage people by reference to their sex, race, age or other such characteristics.

This is the background against which this paper has considered risks associated with the use of AI systems and large data sets to draw insights and make predictions about individuals. While the paper has focused on a particular type of decision making as a case study, the risks identified here apply to almost any commercial context in which predictive modelling is used to assist in decision making, such as financial and insurance services, telecommunications and human resources.

The simulation results in this paper demonstrate that unfair outcomes arising from algorithmic bias may engage the right to equality and non-discrimination for an individual or

group. Depending on the particularities of the situation, this could result in unlawful discrimination under Australian law.

When AI systems produce unfair results, this may sometimes meet the technical legal definition of unlawful discrimination. Regardless of the strict legal position, there is always a strong imperative to identify and address algorithmic bias, especially where unfairness disproportionately affects people who already experience disadvantage. Scenario 1 is a clear example of a disproportionate impact on one population group, where Aboriginal and Torres Strait Islander peoples are less likely to be offered a competitive service contract due to existing societal inequalities reflected in the AI system.

Risks of harm must be considered in the context in which they arise—the consequences of unfair outcomes are more serious when considering equal access to an essential service. Although the electricity retail market is an example for this simulation, it is a useful context to consider how disparate impacts on groups and individuals have everyday consequences. Access to affordable, reliable and sustainable energy is a basic need for everyone in Australia.[58] Consumers will be economically disadvantaged when they are excluded from competitive or cheaper service contracts. As noted previously, these consequences are likely to further entrench existing inequalities. These practices may also result in a redistribution of costs and benefits that reduce overall consumer welfare and social equality.[59]

Service providers across many industries have access to vast amounts of data from data brokers, from which they can make predictions about individuals and population groups.[60] This type of profiling, enabled by increased data collection and analysis, can lead to differential treatment of consumer groups (segmentation practices), such as opaque consumer targeting[61] and price discrimination.[62] The Australian Competition and Consumer Commission (ACCC) describes these types of harm to consumers as 'risks from increased consumer profiling' and 'discrimination and exclusion'.[63]

Segmentation of consumer groups for targeted advertising and offers may protect certain groups, such as children, from inappropriate content. However, these practices can also cause disadvantage, especially where they exclude population groups, based on protected attributes and other characteristics. Such exclusion, especially where it results in unlawful discrimination, can cause practical harm to affected individuals, reduce consumer trust and engagement, limit consumer choice and control, and lessen market competition. People have little or no choice about whether they are subject to these practices. These factors make it difficult or

impossible for consumers to take action to protect themselves.

The simulation highlights the urgent need for businesses to detect and address risks of harm to individuals and groups in the community, and the importance of a regulatory framework that protects the community from these harms.

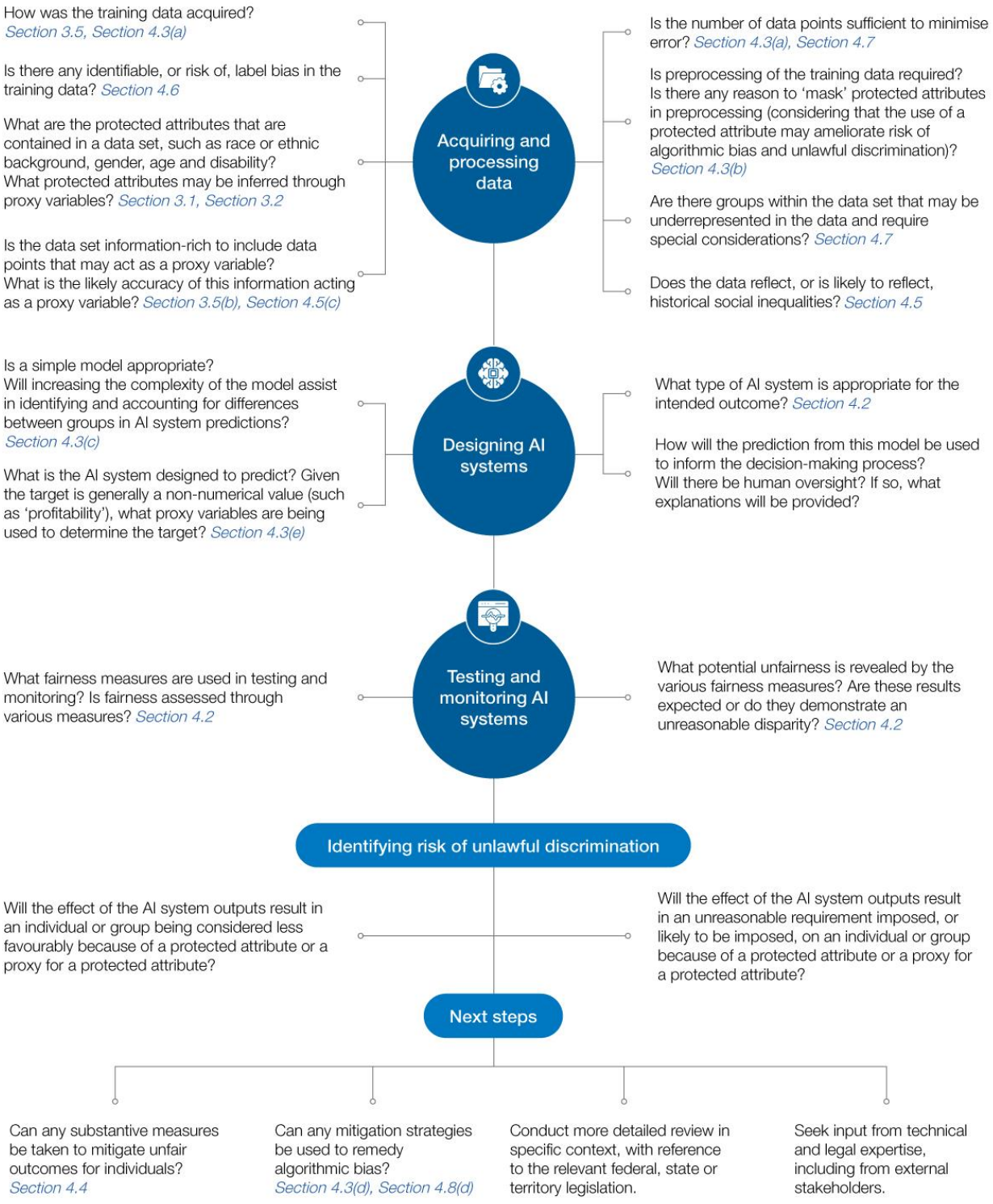
5.2 Protecting the right to equality and non-discrimination

(a) Responsible business use of AI and data

As observed at Section 3.2, it is unlawful to discriminate on the basis of protected attributes and businesses should be proactive in identifying the human rights risks or impacts of their practices.[64] Addressing the problem of algorithmic bias in AI systems reduces the legal and reputational risk of making unfair and inaccurate decisions, some of which could involve unlawful discrimination.

Businesses are seeking guidance on how to integrate human rights protections into their operations, particularly in the context of AI and data. We set out key questions that assist designers to identify 'red flags' when considering algorithmic bias and mitigation strategies. They are based on the simulation scenarios, and not intended to be an exhaustive list of considerations.

Responsible business use of AI and data



(b) Complying with existing laws

Using AI in the sorts of consumer context explored in this paper must be fair and lawful. Our laws are intended to protect people from these kinds of harms and can be used to hold businesses accountable for their practices. As previously observed, our existing laws, which include anti-discrimination laws[65] and Australian Consumer Law,[66] protect the Australian public in a number of ways.

Businesses that use these tools to assist in their decision-making processes must be accountable for their design and deployment. There is a growing focus on improving accountability in the use of AI.[67]

Addressing algorithmic bias in AI systems will improve outcomes for all Australians. Anyone designing or using these AI systems should be alert to erroneous or unjustified differential treatment between groups and take steps to mitigate algorithmic bias.

Several steps are required to challenge the lawfulness of a decision under anti-discrimination legislation—the recipient of a decision would need to understand the basis of the decision, and be able to prove the legal elements of unlawful discrimination in their particular case. As noted at Section 5.1, people are often unaware or have no choice about how decisions are made, which affect them. This, in turn, can make it more difficult to challenge unlawful or otherwise unfair decisions.

(c) Broader reform to promote accountability in the development and use of AI

While this paper focuses on some important issues associated with algorithmic bias, all of the partners in this project are also working on broader questions related to the development and use of AI. In particular, the accountability of AI systems in decision making raises challenging questions of policy and law, which are considered more fully in the Australian Human Rights Commission's major project on Human Rights and Technology.[68]

In addition, the Consumer Policy Research Centre (CPRC) promotes a coordinated approach for AI system regulation in the digital economy, which clearly outlines the *consumer outcomes* companies should seek to uphold in the design and deployment of AI tools.[69] This would help inform businesses of risks of harm and promote innovation consumers can trust,[70] while also providing strong protections for consumers across the economy.[71]

CPRC considers the consumer outcomes, which would promote responsible business use of AI and data, include:	
	***Accessibility*** Markets are inclusive, and all consumers have the right to access this technology and its application on an equal basis with others.
	***Accountability*** Consumers have a clear route for seeking explanations and accessing appropriate redress from a responsible party if things go wrong.
	***Agency*** Consumers are empowered to exercise autonomy and freedom of choice in their interactions with technologies such as AI systems and the use of their personal data.
	***Transparency*** People are made aware when they are the subject of a decision-making process that uses an AI system.
	***Understandability and explainability*** Individuals subject to these decisions are entitled to a meaningful, comprehensible explanation of the AI system and its decision-making process.
	***Sustainability*** Long-term implications of technology on consumers are considered and addressed throughout design and implementation.

The Australian Government is already considering gaps and potential reform in protections frameworks regarding consumers' privacy rights, transactional bargaining power, and consumer choice and control over their data.[72] It is also progressing Consumer Data Right reforms, which intend to increase competition in markets—including electricity—through greater data access and portability.[73] Economy wide principles that promote positive consumer outcomes from AI systems may work alongside these other reforms to improve competition, data privacy and fairness. Together, these would help minimise consumer harms such as discrimination and exclusion and associated negative effects on consumer trust, choice, and control.

Appendices

6 Appendix 1: Glossary

AI system means a machine learning system that makes decisions or predictions for a specific narrow task that would be considered intelligent behaviour if performed by a human (such as deciding whether to offer a customer a product based on their application form). AI systems are often based on a statistical model which they learn from historical data.

Artificial intelligence (AI) is a broad term lacking a definitive definition. An analysis of the literature is beyond the scope of this paper. This paper refers to AI as a machine learning system, defined as an **AI system**.

Base rates means the distribution of feature and target values will differ between groups. For a binary classification we may refer to the fraction of a group who have a positive target label as the group's base rate. For example, we might say that 60% of male applicants are suitable for a job interview.

Binary classification means a modelling problem where the target is the answer to a yes or no question such as "is the individual going to be profitable?"

Cohort means a number of individuals whose data are assessed by the AI system.

Decision (or prediction) means the result of the AI system for each individual. In some AI systems, the result may including taking action for each individual (i.e. automatically accepting or rejecting an application), while others may predict a factor or score (i.e. likelihood of repayment) which may be converted into a decision by additional logic (such as applying an acceptance threshold based on risk appetite). In this simulation, the result is a prediction of whether or not an individual is profitable.

Equality is predicated on the idea that all human beings are born free and equal. It means that all persons are equal before the law and are entitled without any discrimination to the equal protection of the law.

Fairness measures means the mathematical expressions to quantify the fairness of an AI system applied to a particular cohort. There are many reasonable definitions of fairness that relate to equality or equity (such as equal opportunity or selection parity), yet they are often in conflict.

Feature means a known attribute regarded as characteristic of a person or instance, such as their age, income, or postcode. The data set may contain many features.

Formal equality is concerned with equality of treatment and expects all people to be treated the same way regardless of their differences.

Group means a sub-cohort in which all individuals share a common (protected) attribute, such as *male customers*.

Label means the value of the target variable in the training data set for a particular person.

Machine learning algorithms are algorithms that take sample data (known as training data) and a mathematical description of a goal, to produce predictions, decisions or actions as outputs that aim to drive the goal as informed by the training data.

Model means a mathematical representation of a relationship between features and the target.

Outcomes means the effect of the AI system, and is generally discussed in relation to the potential harms and benefits.

Protected attribute means an attribute of a person (including age, disability, race, sex), the basis of which is unlawful to discriminate in certain areas of public life, protected under federal and state anti-discrimination legislation.

Proxy for a variable (proxy variable) means a feature that is distinct from the variable in question but potentially contains some information about it. For instance, postcode may be a proxy for socio-economic status because certain neighbourhoods are wealthier than others.

Selected individuals, in these scenarios, are people that the AI system predicts will be profitable and are thus chosen to receive a market-competitive service contract.

Suitable for selection, in these scenarios, is a quality of someone who, if chosen by the AI system to receive a market-competitive contract, would become a profitable customer. The selection of suitable individuals is therefore deemed correct, or accurate.

Simulated data means fictitious data generated by a simulation using a set of assumptions about the world.

Simulation means an investigative tool to generate synthetic data using a model based on a set of assumptions about the world.

Substantive equality is concerned with equality of opportunity and outcomes. It recognises that formal equality does not address underlying, historical and structural inequalities that limit a person's opportunity to participate equally in society. Substantive equality goes beyond equal treatment, and attempts to redress underlying, historical and structural inequalities, which can require the use of affirmative action or 'special measures'.

Target (or target variable) means the feature that the system is attempting to predict. In this simulation, the target variable is whether or not an individual is profitable.

Training data means data characterising a historical cohort, containing both feature and target variables, used to inform model selection.

Unmeasured features mean a collection of unmeasured features generated using random noise. In this simulation, these include features that may affect profitability not captured by other listed features, such as whether or not the individual tends to pay their bills on time or use power in predominantly off-peak or on-peak times.

7 Appendix 2: Technical details

This appendix describes the mechanisms used to generate the data in each of the simulated scenarios discussed above.

The code used to generate the data, predictions and results in this report can be found here:

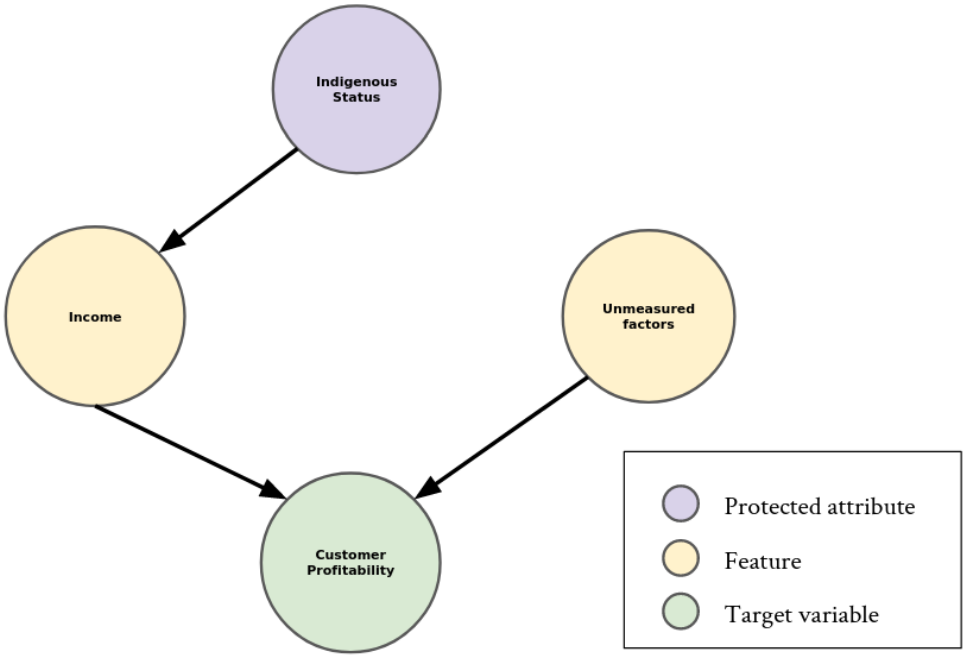
https://github.com/gradientinstitute/sim_algo_bias

These simulation results, tools and methodology are experimental, not to be used as an automated decision-making tool, and are not endorsed by the other partners to this paper.

7.1 Scenario 1

The data sets used in Scenario 1 were generated using the causal relationships illustrated in Figure 7.1.

Figure 7.1 Causal graph of the data generation process for Scenario 1.



Our explanation and assumptions are:

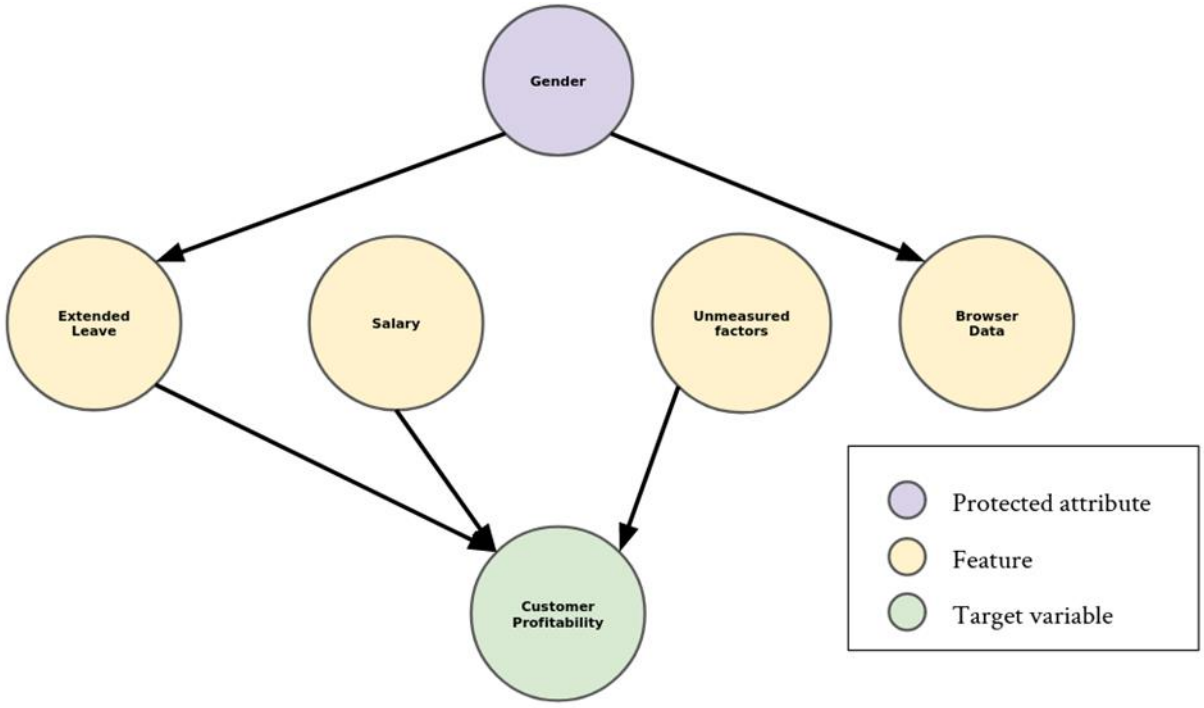
- The protected attribute for Scenario 1 is race.
- An individual's profitability is causally dependent on:
  - income
  - a collection of unmeasured factors or features generated using random noise, including features that may affect profitability not captured by other listed features, such as whether or not the individual tends to pay their bills on time or use power in predominantly off-peak or on-peak times (**unmeasured features**).

- There is a causal relationship between an Aboriginal or Torres Strait Islander person's race and income.
- There is not a causal relationship between an Aboriginal or Torres Strait Islander person's race and the unmeasured features.

7.2 Scenario 2

The data sets used in Scenario 2 were generated using the causal relationship illustrated in Figure 7.2.

Figure 7.2 Causal graph of the data generation process for Scenario 2.



Our explanation and assumptions are:

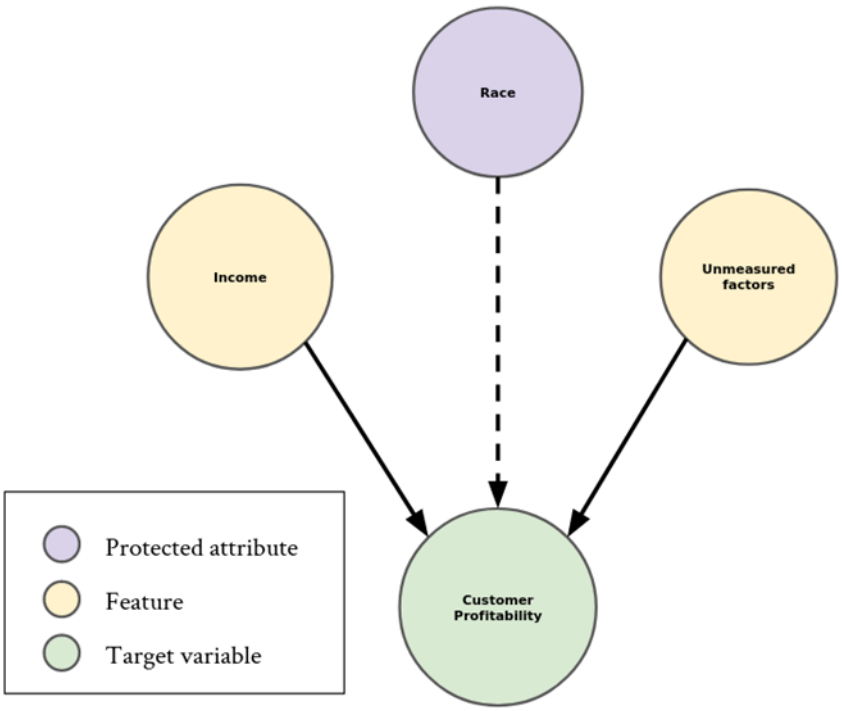
- The protected attribute in Scenario 2 is sex.
- The sex of each person in the data set is randomly assigned with an equal probability of being male or female.
- An individual's profitability is causally dependent on:
  - salary
  - income, based on potential to take extended leave (on average, females have a higher probability of taking an extended leave from the workplace)
  - unmeasured features.

- There is a causal relationship between an individual's sex and:
  - income, based on potential to take extended leave (on average, females have a higher probability of taking an extended leave from the workplace)
  - potential to take extended leave
  - browser data.
- There is not a causal relationship between an individual's sex and:
  - salary
  - unmeasured features.
- The AI system can observe income, browser history, sex (unless specified in the scenario text) and whether the customer is profitable.
- The AI system cannot observe whether an individual will take extended leave from the workplace.

7.3 Scenario 3

The data set used in Scenario 3 was generated using the causal relationship illustrated in Figure 7.3.

Figure 7.3 Causal graph data generation process for Scenario 3.



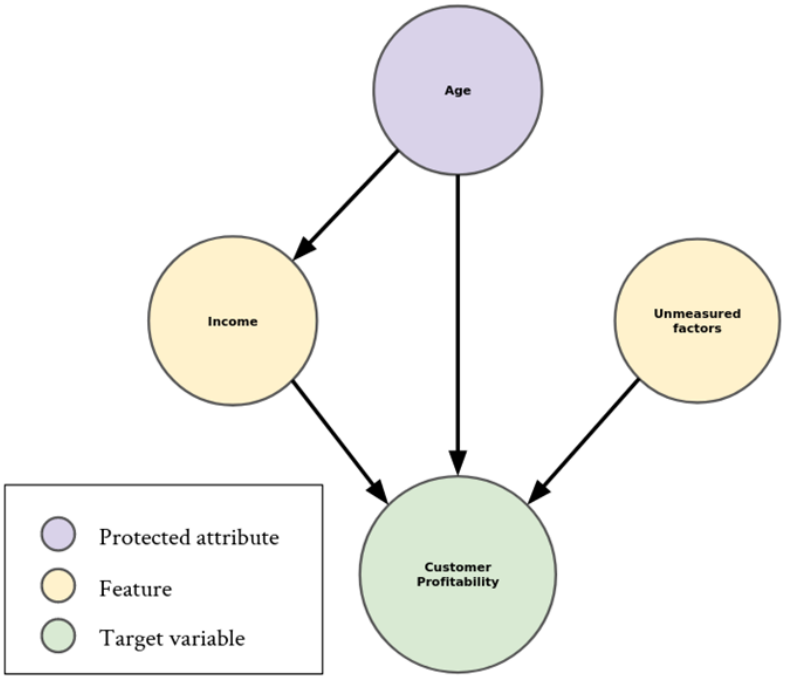
Our explanation and assumptions are:

- The protected attribute in Scenario 3 is race.
- An individual's profitability is causally dependent on:
  - race (related to the label bias introduced by discriminatory behaviour of customer support staff towards Asian Australians which does not reflect the true thing being predicted)
  - unmeasured features.
- There is not a causal relationship between an individual's race and income.

7.4 Scenario 4 and Scenario 5

The data set used in Scenario 4 and 5 were generated using the causal relationship illustrated in Figure 7.4.

Figure 7.4 Causal graph of data generation process for Scenario 4 and Scenario 5.



Our explanation and assumptions are:

- The protected attribute in Scenario 4 and Scenario 5 is age.
- An individual's profitability is causally dependent on:
  - age (identifying different behaviours from different age groups)
  - income
  - unmeasured features.
- There is a causal relationship between an individual's age and income.

8 Appendix 3: Group Accuracies

Table 8.1 provides the accuracy of the AI system's predictions for each group before and after mitigation. Table 8.2 indicates which groups are referred to as the advantaged (Adv) and disadvantaged (Dis) in Table 8.1.

Table 8.1 Accuracies for each group before and after mitigation

	Pre-Mitigation (Adv/Dis)	Post-Mitigation (Adv/Dis)
Scenario 1	0.91 / 0.87	0.91 / 0.79
Scenario 2	0.93 / 0.88	0.93 / 0.93
Scenario 3	0.94 / 0.93	0.94 / 0.94
Scenario 4	0.94 / 0.87	0.95 / 0.90
Scenario 5	0.85 / 0.91	0.86 / 0.94

Table 8.2 The advantaged and disadvantaged groups in each scenario

	Advantaged Group	Disadvantaged Group
Scenario 1	Non-Aboriginal and Torres Strait Islander peoples	Aboriginal and Torres Strait Islander peoples
Scenario 2	Males	Females
Scenario 3	Non-south-east Asian Australians	South-east Asian Australians
Scenario 4	Over 25 year olds	Under 25 year olds
Scenario 5	Over 25 year olds	Under 25 year olds

Endnotes

[1] Terms that appear at first in bold are defined in Appendix 1.

[2] United Nations Office of the High Commissioner for Human Rights, *Guiding Principles on Business and Human* Rights*: Implementing the United Nations "Protect, Respect and Remedy" Framework* (2011) 20. Endorsed by the Human Rights Council on 16 June 2011 in Resolution 17/4.

[3] This form of narrow AI is referred to as "supervised learning" in the machine learning literature. Other categories of machine learning include unsupervised and reinforcement learning. These are susceptible to many of the same forms of algorithmic bias discussed here (and others), however they are not the focus of this report.

[4] Australia is a signatory to seven core human rights treaties, which cover civil and political rights, and economic, social and cultural rights. Accordingly, Australia has voluntarily agreed to comply with human rights standards and to integrate them into domestic law, policy and practice.
To be fully enforceable in Australia, international human rights law must be incorporated into domestic Australian law through legislation, policy and other arrangements. Where international law is incorporated, this creates rights, obligations, and accountability mechanisms *under Australian law*, which apply to individuals, public and private organisations.
Human rights are protected in Australia in various ways, including protections in federal, state and territory anti-discrimination and equal opportunity laws.

[5] *International Covenant on Civil and Political Rights*, opened for signature 16 December 1966, 999 UNTS 171 (entered into force 23 March 1976) arts 2, 26; *Convention on the Elimination of All Forms of Discrimination against Women*, opened for signature 18 December 1979, 189 UNTS 1429 (entered into force 3 September 1981) art 11; *International Convention on the Elimination of All Forms of Racial Discrimination,* opened for signature 21 December 1965, 660 UNTS 195 (entered into force 4 January 1969) arts 2, 5; *Convention on the Rights of Persons with Disabilities,* opened for signature 13 December 2006, 2515 UNTS 3 (entered into force 3 May 2008) art 27.

[6] *Australian Human Rights Commission Act 1986* (Cth); *Racial Discrimination Act 1975* (Cth); *Sex* Discrimination *Act 1984* (Cth); *Disability Discrimination Act 1992* (Cth); *Age Discrimination Act 2004* (Cth); *Fair Work Act 2009* (Cth); see also more general human rights legislation in Victoria, Queensland and the Australian Capital Territory.

[7] See e.g. *Universal Declaration of Human Rights*, GA Res 217A (III), UN GAOR, 3rd sess, 183rd plen mtg, UN Doc A/810 (10 December 1948) arts 1, 7; *International Covenant on Civil and Political Rights*, opened for signature 16 December 1966, 999 UNTS 171 (entered into force 23 March 1976) art 26.

[8] *Australian* Human *Rights Commission Act 1986* (Cth); *Racial Discrimination Act 1975* (Cth); *Sex Discrimination Act 1984* (Cth); *Disability Discrimination Act 1992* (Cth); *Age Discrimination Act 2004* (Cth).

[9] See e.g. *Racial Discrimination Act 1975* (Cth) s 13; *Sex Discrimination Act 1984* (Cth) s 22; *Age Discrimination Act 2004* (Cth) s 28; *Disability Discrimination Act 1992* (Cth) s 24.

[10] See, for example: *Discrimination Act 1991* (ACT) and *Anti-Discrimination Act 1991* (Qld).

[11] See Australian Competition Consumer Commission, *Legislation* (Web Page) <https://www.accc.gov.au/about-us/australian-competition-consumer-commission/legislation>.

[12] See Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 *California Law Review* 671; Ansgar Koene, 'Algorithmic Bias: Addressing Growing Concerns' (June 2017) *IEEE Technology and Society Magazine,* 31 <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7947257>; Nicol Turner Lee, Paul Resnick, and Genie Barton, 'Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms' *Brookings* (online, 22 May 2019) <https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/>.

[13] The more general concept of 'bias' in law generally refers to the natural justice principle that a judge or other decision maker should be free from actual bias or the appearance of bias. This principle is intended to avoid a deviation from the true course of decision making and to maintain public confidence in the administration of justice. See *Johnson v Johnson* (2000) 201 CLR 488, 501 (Kirby J); *Ebner v Official Trustee in Bankruptcy* (2000) 205 CLR 337; *Minister for Immigration and Multicultural Affairs v Jia Legeng* (2001) 205 CLR 507. For the avoidance of doubt, we do not use the term bias in this legal sense in this paper.

[14] Lawful exemptions to differential treatment based on statistical and actuarial data exist in the provision of insurance and superannuation in certain circumstances. See, for example: *Disability Discrimination Act 1992* (Cth) s 46.

[15] Brigid Richmond and Consumer Policy Research Centre, *A Day in the Life of Data: Removing the opacity* surrounding *data collection, sharing and use environment in Australia* (Report, May 2019) 6–12 <www.cprc.org.au/wp-content/uploads/CPRC-Research-Report_A-Day-in-the-Life-of-Data_final-full-report.pdf>; Phuong Nguyen and Lauren Solomon, 'Consumer data and the digital economy: Emerging issues in data collection, use and sharing' *Consumer Policy Research Centre* (Report, 17 July 2018) <https://cprc.org.au/publication/consumer-data-and-the-digital-economy-report/>.

[16] The ACCC considers that the prevalence of data collection and use, by both digital platforms and other businesses, and the resulting increased potential for significant harm to consumers, demands consideration of the current protections provided to consumers. See: Australian Competition and Consumer Commission, *Digital Platforms Inquiry* (Final Report, June 2019) 498 <https://www.accc.gov.au/system/files/Digital%20platforms%20inquiry%20-%20final%20report.pdf>.

[17] The Australian Government accepted the ACCC recommendation to review aspects of the *Privacy Act 1988* (Cth), and will conduct further consultation on the recommendation to ensure that the definition of 'personal information' captures technical data and other online identifiers that raises privacy concerns: Australian Government, *Regulating in the digital age: Government response and implementation roadmap for the Digital Platforms Inquiry* (December 2019) 17 <https://treasury.gov.au/sites/default/files/2019-12/Government-Response-p2019-41708.pdf>; see also *Privacy Act 1988* (Cth); Office of the Australian Information Commissioner, *Guide to data analytics and the Australian Privacy Principles* (March 2018) <https://www.oaic.gov.au/privacy/guidance-and-advice/guide-to-data-analytics-and-the-australian-privacy-principles/>.

[18] Shoshana Zuboff, 'Big Other: Surveillance Capitalism and the Prospects of an Information Civilisation' (2015) 30(1) *Journal of Information Technology* 75, 85; Brigid Richmond and Consumer Policy Research Centre, *A Day in the Life of Data: Removing the opacity surrounding data collection, sharing and use environment in Australia* (Report, May 2019) 6–12 <www.cprc.org.au/wp-content/uploads/CPRC-Research-Report_A-Day-in-the-Life-of-Data_final-full-report.pdf>; Phuong Nguyen and Lauren Solomon, 'Consumer data and the digital economy: Emerging issues in data collection, use and sharing' *Consumer Policy Research Centre* (Report, 17 July 2018) <https://cprc.org.au/publication/consumer-data-and-the-digital-economy-report/>; citing Wolfie Christl and Sarah Spiekermann, *Networks of Control: A Report on Corporate Surveillance, Digital Tracking, Big Data & Privacy* (Facultas Verlags- und Buchhandels AG, 2016).

[19] Equifax, *Analytic Dataset* (Web Page) <https://www.equifax.com/business/analytic-dataset/>; Experian, *ConsumerView: Marketing data that connects brands with fans* (Web Page) <https://www.experian.com/marketing-services/targeting/data-driven-marketing/consumer-view-data>.

[20] CHOICE contacted several companies. The following companies engaged in discussion or provided a quote for purchase of data sets: Equifax (The Prospect Shop), Experian, RDA Research, Impact Lists and Funnel. Not all organisations provided a response, and those organisations that did provided varying quotes.

[21] Australian Competition and Consumer Commission, *Restoring electricity affordability and Australia's* competitive *advantage* (Retail Electricity Pricing Inquiry—Final Report, June 2018) 247; Brendon O'Neill, 'Energy Credit Checks Explained' *Canstar* (online, 4 December 2019) <https://www.canstarblue.com.au/electricity/energy-credit-checks-explained/>.

[22] Kara Brandeisky, *Online data brokers are being increasingly employed by third-party companies to improve the accuracy of their predictive algorithms* (online, 5 June 2014) <https://money.com/data-brokers-online-privacy-tools/>.

[23] Except in specific circumstances which commonly relate to consumers under a "deemed customer retail arrangement" provided for in the National Electricity Market regions of New South Wales, Victoria, South Australia, Tasmania and South East Queensland. See Australian Energy Market Commission, *National Energy Retail Rules Version 20* (19 March 2020) 44 <https://www.aemc.gov.au/sites/default/files/2020-03/NERR%20v20%20full.pdf>; Essential Services Commission Victoria, *Energy Retail Code Version 15* (2 February 2020) 60 <https://www.esc.vic.gov.au/sites/default/files/documents/energy-retail-code-v15-pdf-version.pdf>.

[24] Unless otherwise specified, we use 10,000 points to train the model. Simulation results were the same when tested with 20,000 and 50,000 data points. The simulation took longer to run when using the larger data sets.

[25] We used the logistic regressor classifier from the popular Python-based machine learning library Scikit Learn as the predictive model in our AI system for this case study; see Christopher Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006) 205.

[26] More complex types of AI systems may present different challenges when attempting to identify and mitigate algorithmic bias, such as interrogating black box systems and uncovering (unwanted) correlations between features.

[27] It is more difficult to ascertain the importance of features in systems that use more complex models due to their non-linear nature. Some common approaches used in practice include partial dependence plots, individual conditional expectation and local surrogates. See, for example: <https://christophm.github.io/interpretable-ml-book/feature-importance.html>.

[28] Numerous group fairness measures have been proposed in the AI literature. See <https://shubhamjain0594.github.io/post/tlds-arvind-fairness-definitions/>. To a first approximation, they fall into three main categories. See <https://fairmlbook.org/classification.html>. The three fairness measures selected for this simulation each belong to a different category, to efficiently cover the breadth of possible fairness measures.

[29] See Alessandro Mantelero, 'Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection' (2016) 32(3) *Computer Law & Security Review* 238, 238; Virginia Eubanks, *Automating Inequality: How high tech tools profile, police and punish the poor* (St Martin's Press, 2017) 6, 7.

[30] Australian Capital Territory law prohibits discrimination on the basis of accommodation status and employment status in certain circumstances: *Discrimination Act 1991* (ACT).

[31] European Parliament, A governance framework for algorithmic accountability and transparency (Framework, PE 624.262, April 2019) <https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf>; Google, Responsible AI Practices (Web Page, 2020) <https://ai.google/responsibilities/responsible-ai-practices/>.

[32] Rich Zemel et al, 'Learning Fair Representations' (Conference Paper, International Conference on Machine Learning, 17 June 2013) 325.

[33] Benjamin Tayo, 'Simplicity versus Complexity in Machine Learning', *Towards Data Science* (Blog Post, 11 November 2019).

[34] Christopher Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006) 6.

[35] Brian d'Alessandro, Cathy O'Neil and Tom LaGatta, 'Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification' (2017) 5(2) *Big Data* 120.
[36] Robert Williamson and Aditya Krishna Menon, 'Fairness Risk Measures' (Conference Paper, International Conference on Machine Learning, 9 June 2019) 6786.
[37] Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan, 'Inherent trade-offs in the fair determination of risk scores' (Conference Paper, Innovations in Theoretical Computer Science, 2017).
[38] Australian Government, *Closing the Gap Report 2020* (Closing the Gap Report No 12, 12 February 2020) <ctgreport.niaa.gov.au>.
[39] Megan Weier et al, *Money Stories: Financial resilience among Aboriginal and Torres Strait Islander Australians* (Report, May 2019).
[40] See Australian Human Rights Commission, *Gender Equality* (Web Page) <https://humanrights.gov.au/quick-guide/12038>; Australian Human Rights Commission, *Face the facts: Gender Equality* (Report, 2018) <https://humanrights.gov.au/our-work/education/face-facts-gender-equality-2018>.
[41] See, for example, Productivity Commission, *Paid Parental Leave: Support for Parents with Newborn Children* (Inquiry Report No 47, 28 February 2009); Productivity Commission, *Childcare and Early Childhood Learning: Overview and Recommendations* (Inquiry Report No 73, 31 October 2014).
[42] See discussion as Section 1.1 (a); for example, Pinterest users are 71% female (Jessica Clement, 'Distribution of Pinterest users worldwide as of July 2020, by gender', *Statista* (Web Page, 24 July 2020) <https://www.statista.com/statistics/248168/gender-distribution-of-pinterest-users/>). Additionally, academic studies have shown that gender prediction from web-browsing data can achieve around 80% accuracy (Do Viet Phuong and Tu Minh Phuong, 'Gender Prediction Using Browsing History' in *Knowledge and Systems Engineering* (Advances in Intelligent Systems and Computing, Springer, 2014) 271-283).
[43] Kara Brandeisky, *Online data brokers are being increasingly employed by third-party companies to improve the accuracy of their predictive algorithms* (Web Page, 5 June 2014) <https://money.com/data-brokers-online-privacy-tools/>.
[44] This behaviour is the algorithmic equivalent of "Redlining", a historical practice in the United States where public and private services were denied to residents of specific neighbourhoods, disproportionately affecting minorities. See Cynthia Dwork et al, 'Fairness through Awareness' (Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, January 2012) 214-226; Amy Hillier, 'Redlining and the Home Owners' Loan Corporation' (2003) 29(4) *Journal of Urban History* 394.
[45] For further discussion on the effects of unfairness through redundant feature encoding, see Alistair Reid and Simon O'Callaghan, *Ignorance Isn't Bliss* (Web Page, 2019) <https://gradientinstitute.org/blog/2>.
[46] Joao Gama et al, A Survey on Concept Drift Adaptation (2013) 1(1) *ACM Computing Surveys* <http://eprints.bournemouth.ac.uk/22491/1/ACM%20computing%20surveys.pdf>.
[47] Cynthia Dwork et al, 'Fairness through Awareness' (Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, January 2012) 214-226.
[48] *Sex Discrimination Act 1984* (Cth) s 7D.
[49] See Australian Human Rights Commission, 'Unconscious Bias', *It Stops With Me* (Web Page) <https://itstopswithme.humanrights.gov.au/unconscious-bias>.
[50] The Australian Human Rights Commission's consultation with community groups and individuals reveals that people from China and people with south-east Asian backgrounds report an increase in racist treatment throughout the COVID-19 pandemic. This treatment has been reported by some people of ethnic Chinese background from south-east Asian nations such as Malaysia and Singapore but also more broadly by some people from other backgrounds including people from Vietnam and

Indonesia. See also Asian Australian Alliance and Osmond Chiu, *COVID-19 Coronavirus Racism Incident Report* (2020) <https://asianaustralianalliance.net/covid-19-coronavirus-racism-incident-report/covid-19-racism-incident-report-preliminary-report/>; Australian Human Rights Commission, *Racism undermines Covid-19 Response* (Web Page, 8 April 2020) <https://humanrights.gov.au/about/news/racism-undermines-covid-19-response>; Jason Fang, Samuel Yang and Christina Zhou, 'Australians urged to 'show kindness' amid reports of COVID-19 racial discrimination complaints' *ABC News* (online, 3 April 2020) <https://www.abc.net.au/news/2020-04-03/racism-covid-19-coronavirus-outbreak-commissioner-discrimination/12117738>.

[51] Alexandra Feldberg and Tami Kim, 'Beyond Starbucks: How Racism Shapes Customer Service', *New York Times* (20 April 2018); Alexandra Feldberg and Tami Kim, *How Companies Can Identify Racial and Gender Bias in Their Customer Service* (Harvard Business Review, May 2018) <https://hbr.org/2018/05/how-companies-can-identify-racial-and-gender-bias-in-their-customer-service>. This study was based on customer service support via written content (email contact) which included a personal name with strong race and gender associations.

[52] See for example, Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 *California Law Review* 671, 712.

[53] Heinrich Jiang and Ofir Nachum, 'Identifying and Correcting Label Bias in Machine Learning' (Proceedings of the International Conference on Artificial Intelligence and Statistics, 2020) <http://proceedings.mlr.press/v108/jiang20a/jiang20a.pdf>.

[54] For example, arrest data for drug possession contains label bias if minorities are stopped and searched at higher rates than the general population. It is possible to partially quantify this bias by comparing arrest rates for possession across communities with data on drug use for the same communities collected from sewage data - which measures the overall use of drugs within that community.

[55] Typical losses such as mean squared error (**MSE**) and mean absolute error (**MAE**) fall in this category.

[56] In fact, there is a whole space of possible models from those that ignore group status entirely, to that allow the model to vary slightly between groups to completely separate models for each group.

[57] *Age Discrimination Act 2004* (Cth) s 33(c).

[58] UN General Assembly, *Transforming our world: the 2030 Agenda for Sustainable Development* (21 October 2015) A/RES/70/1, 14 [Goal 7].

[59] When price discrimination occurs in essential service markets the economic benefits it may offer (namely improved accessibility to quality products, lower prices on average and stronger competition) must be considered alongside distributive fairness concerns regarding who is paying more for an essential service that they must consume. See: *Financial Conduct Authority, Price discrimination in financial services: How should we deal with questions of fairness?* (Research Note, July 2018) 4-7 <https://www.fca.org.uk/publication/research/price_discrimination_in_financial_services.pdf>.

[60] Australian Competition and Consumer Commission, *Digital Platforms Inquiry* (Final Report, June 2019) 449.

[61] CPRC research highlights examples of how AI system decisions can be based on incorrect assumptions, use poor quality input data, or be inherently biased, leading to discrimination and exclusion of consumers. See: Phuong Nguyen and Lauren Solomon, 'Consumer data and the digital economy: Emerging issues in data collection, use and sharing,' *Consumer Policy Research Centre* (Report, 17 July 2018) 46 <https://cprc.org.au/publication/consumer-data-and-the-digital-economy-report/>.

[62] Other risks and harms identified by the ACCC included: Eroding consumer trust in data-based innovations; consumer profiling and manipulation, loss of consumer autonomy; price discrimination; decreased welfare from reduced competition; decreased consumer privacy; and increased risks for

vulnerable consumers. See Australian Competition and Consumer Commission, *Digital Platforms Inquiry* (Final Report, June 2019) 442–448.

[63] Australian Competition and Consumer Commission, *Digital Platforms Inquiry* (Final Report, June 2019) 373–501. The ACCC describes some of these consumer harms as 'discrimination and exclusion' in the consumer context of the provision of goods and services to the public.

[64] United Nations Office of the High Commissioner for Human Rights, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework* (2011) 20. Endorsed by the Human Rights Council on 16 June 2011 in Resolution 17/4.

[65] *Australian Human Rights Commission Act 1986* (Cth); *Racial Discrimination Act 1975* (Cth); *Sex Discrimination Act 1984* (Cth); *Disability Discrimination Act 1992* (Cth); *Age Discrimination Act 2004* (Cth); *Fair Work Act 2009* (Cth); see also more general human rights legislation in Victoria, Queensland and the Australian Capital Territory.

[66] See Australian Competition Consumer Commission, *Legislation* (Web Page) <https://www.accc.gov.au/about-us/australian-competition-consumer-commission/legislation>.

[67] See, for example: Australian Government, Department of Industry, Innovation and Science, AI Ethics Principles (November 2019) <https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework>; Microsoft, 'Microsoft AI Principles' (Webpage) <https://www.microsoft.com/en-us/AI/our-approach-to-ai>.

[68] Australian Human Rights Commission, *Human Rights and Technology Discussion Paper* (Report, 2019) 91.

[69] For a discussion on outcomes-based regulation see: Australia Communications and Media Authority, *Artificial intelligence in communications and media* (Occasional Paper, July 2020) 46 <https://www.acma.gov.au/sites/default/files/2020-07/Artificial%20intelligence%20in%20media%20and%20communications_Occasional%20paper.pdf>.

[70] The UK Centre for Data Ethics and Innovation has considered consumer trust and innovation, and note that 'in the absence of trust, consumers are unlikely to use new technologies or share the data needed to build them, while industry will be unwilling to engage in new innovation programmes for fear of meeting opposition.' See: UK Centre for Data Ethics, *An overview of the CDEI's AI Barometer* (Blog, 2020) <https://cdei.blog.gov.uk/2020/06/18/overview-cdei-ai-barometer/>.

[71] An economy-wide approach regarding the access and use of consumer data has been advocated for in other international jurisdictions, with a recent example being the European Strategy for Data which suggests that "a cross-sectoral governance framework for data access and use is necessary to create an "overarching framework for the data-agile economy, thereby avoiding harmful fragmentation of the internal market through inconsistent actions." See European Commission, *A European strategy for data* (Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and the Committee of the Regions, 19 February 2020) <www.ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf>. The importance of building community trust is outlined in European Commission, *On Artificial Intelligence A European approach to excellence and trust* (White Paper, 19 February 2020) <https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf>.

[72] Australian Government, *Regulating in the digital age: Government response and implementation roadmap for the Digital Platforms Inquiry* (Report, 2019) <https://treasury.gov.au/publication/p2019-41708>.

[73] Australian Competition and Consumer Commission, *Consumer Data Right: Project Overview* (Web Page) <www.accc.gov.au/focus-area/consumer-data-right-cdr-0>.


AI