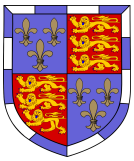
UNIVERSITY OF CAMBRIDGE

Context-conscious fairness throughout the machine learning lifecycle

(Michelle) Seng Ah Lee



St. John's College

Dissertation submitted in September 2022 for the degree of Doctor of Philosophy in Computer Science and Technology

Abstract

As machine learning (ML) algorithms are increasingly used to inform decisions across domains, there has been a proliferation of literature seeking to define "fairness" narrowly as an error to be "fixed" and to quantify it as an algorithm's deviation from a formalised metric of equality. Dozens of notions of fairness have been proposed, many of which are both mathematically incompatible and morally irreconcilable with one another. There is little consensus on how to define, test for, and mitigate unfair algorithmic bias.

One key obstacle is the disparity between academic theory and practical and contextual applicability. The unambiguous formalisation of fairness in a technical solution is at odds with the contextualised needs in practice. The notion of algorithmic fairness lies at the intersection of multiple domains, including non-discrimination law, statistics, welfare economics, philosophical ethics, and computer science. Literature on algorithmic fairness has predominantly been published in computer science, and while it has been shifting to consider contextual implications, many approaches crystallised into open source toolkits are tackling a narrowly defined technical challenge.

The objective of my PhD thesis is to address this gap between theory and practice in computer science by presenting context-conscious methodologies throughout ML development lifecycles. The core chapters are organised by each phase: design, build, test, and monitor. In the design phase, we propose a systematic way of defining fairness by understanding the key ethical and practical trade-offs. In the test phase, we introduce methods to identify and measure risks of unintended biases. In the deploy phase, we identify appropriate mitigation strategies depending on the source of unfairness. Finally, in the monitor phase, we formalise methods for monitoring fairness and adjusting the ML model appropriately to any changes in assumptions and input data.

The primary contribution of my thesis is methodological, including improving our understanding of limitations of current approaches and proposal of new tools and interventions. It shifts the conversation in academia away from axiomatic, unambiguous formalisations of fairness towards a more context-conscious, holistic approach that covers the end-to-end ML development lifecycle. This thesis aims to provide end-to-end coverage in guidance for industry practitioners, regulators, and academics on how fairness can be considered and enforced in practice.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 60 000 words.

(Michelle) Seng Ah Lee
September 2022

Acknowledgements

I would like to express my sincere gratitude to: my supervisor Dr. Jatinder Singh for all his guidance and feedback throughout the past three years, Aviva for my PhD funding, the Alan Turing Institute for the enrichment programme, and my family – my father Jun Bae Lee, my brother Andrew Seng Youb Lee, and my partner Ieuan Reece – for their support and encouragement.

Contents

1 Introduction **16**

1.1 Definitions of AI, ML, and related concepts 17

1.2 Key gaps in academic literature and practical challenges 19

1.2.1 Gap between the narrow algorithmic definitions and the inherent complexity, context-specificity, and subjectivity of "fairness" 19

1.2.2 Gap between the existing tools and the practitioners' requirements for end-to-end, context-specific solutions 21

1.2.3 Gap between the unambiguous fairness tests and real-world uncertainties 22

1.3 Our contribution 22

2 ML Design: contextually defining algorithmic fairness **25**

2.1 Defining key terms 26

2.1.1 Ethics in AI 26

2.1.2 Justice, equality, and equity 27

2.1.3 Discrimination and protected characteristics 28

2.1.4 Bias 28

2.2 Fairness: definitions in computer science 29

2.2.1 Use case: Defining fairness in mortgage lending 30

2.2.2 Flawed assumption: simplicity of separating acceptable inequalities from the unacceptable 33

2.2.2.1 Why fairness cannot always be assessed through legally protected characteristics 35

2.2.2.2 Why fairness cannot always be assessed through relevance vs. irrelevance 35

2.2.2.3 Why fairness cannot always be assessed through effort vs. circumstances 36

2.2.2.4 Why fairness cannot always be assessed through the source of inequality 37

2.2.2.5 The challenge of defining fairness 37

2.3 Lessons from ethical philosophy on (in)equalities . . . 38
2.3.1 Ethical subjectivity of algorithmic fairness . . . 40
2.3.2 Linking ethical philosophy to algorithmic fairness . . . 41
2.4 Lessons from welfare economics: Consideration of welfare and liberty . . . 44
2.4.1 Welfare in algorithmic ethics: beneficence and non-maleficence . . . 45
2.4.2 Liberty in algorithmic ethics: autonomy and explicability . . . 46
2.4.2.1 Autonomy: Liberty . . . 47
2.4.2.2 Autonomy: Forgiveness . . . 47
2.4.2.3 Autonomy: Vulnerability . . . 48
2.4.2.4 Explicability . . . 49
2.5 Proposed method: Key Ethics Indicators . . . 50
2.5.1 Define success . . . 52
2.5.1.1 Special focus: impact on fundamental human rights and vulnerable populations . . . 52
2.5.2 Identify sources of inequality . . . 53
2.5.3 Identify sources of bias . . . 54
2.5.4 Design mitigation strategies . . . 54
2.5.5 Operationalise KEIs and calculate trade-offs between KEIs . . . 56
2.5.6 Select a model and provide justifications . . . 58
2.6 Key chapter takeaways . . . 58

3 ML Build: identifying and measuring unintended unfair biases 60
3.1 How a system's definition as AI and ADM provides only partial information about its risk . . . 61
3.2 Unintended bias risks in ML build lifecycle . . . 65
3.3 Proposed tool: Risk identification questionnaire . . . 67
3.4 Real-world case study: Questionnaire applied to insurance fraud . . . 69
3.4.1 (A) Background information . . . 70
3.4.2 (B) Design: historical/external bias . . . 71
3.4.3 (C) Data collection: representation bias . . . 71
3.4.4 (D) Feature engineering: measurement bias . . . 72
3.4.5 (E) Model build and training: aggregation bias . . . 72
3.4.6 (F) Model evaluation: evaluation bias . . . 73
3.4.7 (G) Model productionisation and monitoring: deployment bias . . . 73
3.5 Survey results on questionnaire's effectiveness . . . 73
3.5.1 Uncovering unexpected biases . . . 74
3.5.2 Ease of integration into existing processes . . . 75
3.5.3 Facilitating communication with non-technical stakeholders . . . 76

3.5.4 Usability and the challenge of designing a simple tool for a challenging problem . . . 76
3.5.5 Helping identify potential mitigation . . . 77
3.6 Illustrating the challenges of identifying a mitigation strategy: Measurement bias due to proxies . . . 78
3.6.1 Legality of proxies . . . 78
3.6.2 Proxies: to include or remove? . . . 79
3.7 Discussion . . . 81
3.7.1 Limitations of this study . . . 81
3.7.2 Future work . . . 82
3.8 Key chapter takeaways . . . 84

4 ML Test: fairness toolkits and trade-off analyses to select the model for deployment 85
4.1 Landscape and gaps in open source fairness toolkits . . . 86
4.1.1 Motivation: lack of guidance on toolkit selection . . . 86
4.1.2 Related work: past studies on fairness challenges in practice . . . 87
4.1.3 Our contribution on fairness toolkits . . . 89
4.1.4 Methodology . . . 90
4.1.5 Open source fairness toolkit feature comparison . . . 92
4.1.6 Key findings from focus group, interviews, and surveys . . . 95
4.1.7 User-friendliness . . . 96
4.1.8 Gaps in toolkit features compared to practitioner needs . . . 102
4.1.9 Contextualisation . . . 103
4.1.10 Discussion: implications and limitations of our study . . . 105
4.1.11 Key takeaways on the landscape and gaps in open source fairness toolkits . . . 109
4.2 Mitigation strategies: critique of "de-biasing" methods . . . 109
4.2.1 Case study: Mitigation strategies for biases in insurance fraud model 111
4.3 Revisiting the Key Ethics Indicators (KEIs) defined in the design phase . . . 113
4.4 Key chapter takeaways . . . 115

5 ML Monitor: risk factors, reviewability, and fairness under uncertainty 116
5.1 Risk factors in AI and automated decision-making . . . 117
5.2 Reviewability: Logging and reporting throughout the lifecycle . . . 119
5.3 Fairness under uncertainty . . . 121
5.3.1 Legal risks of RL: non-discrimination and equality . . . 122
5.3.2 Towards fairness under uncertainty for RL . . . 123
5.3.3 Related work in uncertainty and RL . . . 124

5.3.4 Taxonomy of uncertainty 125
5.3.5 Discussion and future work on fairness under uncertainty 128
5.4 Key chapter takeaways 129

6 Conclusion **131**

A Bias in Model Development Lifecycle Questionnaire **152**

Index **158**

Glossary

algorithm computational method, formula, or procedure [**Lee** et al., 2022].

artificial intelligence (AI) "the designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment" [Russell and Norvig, 2002].

automated decision-making (ADM) decision-making process that involves a substantial automated component by technological means. Solely ADM has no human involvement ([European Commission, c], p. 6; [European Union] Article 22 and Recital 71) to be distinguished from non-solely (or partial) ADM.

bias a colloquial reference to prejudice against one person or group, especially in a way which could be considered to be unfair [**Lee** et al., 2021]. This is often a confusing term in computer science, as "bias" in machine learning, statistics, and econometrics refers to a type of error in a learning algorithm that results in under-fitting the data (compared to *variance*), which results from using too simple of a model to represent a complex relationship [Hastie et al., 2009; Mayson, 2018]. It is used in cognitive sciences as a systematic error in thinking when people are processing information, which are evolutionary functions that form "shortcuts" for the human mind [Haselton et al., 2015]. In clinical research, "statistical bias" refers to the systematic difference between results and facts that is introduced in the design or conduction of research [Tripepi et al., 2008]. For the purpose of this thesis, we define bias as unintended and potentially harmful skewing of algorithmic predictions.

data processing any operation or set of operations which is performed on personal data [European Union].

direct discrimination "direct" discrimination concerns differential or disparate treatment based on a protected characteristic [Wachter et al., 2020].

discrimination From a legal standpoint, *discrimination* refers to the notion that *protected characteristics* should not result in a relative disadvantage of deprivation.

equality a state in which no one is worse off than one another, a quality desired among egalitarians [Fleurbaey, 2015].

equity Reduction of avoidable inequalities, such as the absence of systematic disparities in health between social groups who have different levels of underlying social advantage/disadvantage [Braveman and Gruskin, 2003].

ethics Systematic conceptualisation of 'right' and 'wrong' behaviour, which are often reflected in an accepted set of rules and principles. Over 160 guidelines related to data and AI ethics have been proposed globally, while various organisations have selected combinations of principles into the multitude of "AI ethics" frameworks [AlgorithmWatch, 2019]. Five common themes have been identified across these sets of principles: beneficence, non-maleficence, autonomy, justice, and explicability [Floridi and Cowls, 2022].

fairness toolkits Fairness toolkits are pre-packaged code, often with a user interface, that takes a data set and/or a pre-trained model as the input. The users specify which outcome and predictions they want to test for bias and against which sensitive feature. The outputs are the fairness test results and accompanying visualisations. Some toolkits offer the ability to "de-bias", including pre-processing, in-processing, and post-processing methods. The intent is to inform the developers in assessing whether a model is fair so that this can in turn inform real-world decisions [**Lee** and Singh, 2021a]. Examples include: [Bellamy et al., 2019; Saleiro et al., 2018b; Microsoft and contributors, 2019].

General Data Protection Regulation (GDPR) The General Data Protection Regulation is a regulation in EU law on data protection and privacy in the European Union and the European Economic Area [European Union].

in-processing "de-biasing" techniques proposed for building an algorithm with bias-related constraints, including *adversarial de-biasing* that maximises accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions [Zhang et al., 2018] and *prejudice remover* that adds a discrimination-aware regularisation term to the learning objective [Kamishima et al., 2012] .

indirect discrimination "indirect" discrimination represents an inadvertent negative impact on a protected group [Wachter et al., 2020].

justice Justice is defined here in accordance with legal and organisational science literature, with justice denoting adherence to the standards agreed upon in society (for example,

based on laws) and fairness as a related principle of an evaluative judgement of whether a decision is morally right [Goldman and Cropanzano, 2015].

machine learning (ML) A set of techniques used in AI to detect and extrapolate patterns from data, including supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning. ML is used to adapt to new circumstances and to detect and extrapolate patterns.

model a formal, usually quantitative, representation of a real-life phenomenon by which a prediction, decision, or recommended action is derived, given known factors and assumptions. A model can be based on data and/or expert knowledge, by humans and/or by automated tools like machine learning algorithms [**Lee** et al., 2022].

post-processing "de-biasing" techniques proposed for adjusting the output predictions of an algorithm, including *equalised odds post-processing* that changes output labels to optimise equalised odds [Hardt et al., 2016] and *reject option classification* gives favourable outcomes to unprivileged groups and vice versa around the decision boundary [Kamiran et al., 2012] .

pre-processing "de-biasing" techniques proposed for removing bias from the data before the algorithm build, including generating weights for training samples [Kamiran et al., 2012], learning a probabilistic transformation that edits the features and labels in the data [Calmon et al., 2017], finds a latent representation that encodes the data well but obfuscates information about protected attributes [Zemel et al., 2013], and editing feature values to increase group fairness while preserving rank-ordering [Feldman et al., 2015].

process a series of logical / ordered operations involved in decision-making, encompassing both the technical and business actions taken [**Lee** et al., 2022].

profiling automated personal data processing with the objective of evaluating personal aspects about a natural person, including to analyse or predict aspects of their performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.([European Union], Article 4(4); [European Union], Recital 71; [European Commission, c], p. 6.).

protected or sensitive characteristics *Protected* or *sensitive* characteristics are those commonly referenced and reflected in non-discrimination laws, such as race and ethnicity, gender, religion, age, disability, and sexual orientation, given these personal demographic features are central to discussions on algorithmic fairness.

proxies features that encode information about **protected or sensitive characteristics**.

reinforcement learning the agent learns from a series of reinforcements—rewards or punishments. [Russell and Norvig, 2002].

robustness "The degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions" [rob, 1990].

semi-supervised learning Semi-supervised learning is when an algorithm is given a few labelled examples and must make what we can of a large collection of unlabelled examples [Russell and Norvig, 2002].

supervised learning In supervised learning the agent observes some example input–output pairs and learns a function that maps from input to output [Russell and Norvig, 2002].

system a set of interacting data, algorithm(s), and/or model(s) to form a technical workflow or product, e.g. a facial recognition algorithm that triggers an identity verification model [**Lee** et al., 2022].

technique method used in the technical design, build, and testing of the algorithm [**Lee** et al., 2022].

unintended biases For the purpose of this thesis, we define bias as unintended and potentially harmful skewing of algorithmic predictions, with six categories: historical bias, representation bias, measurement bias, aggregation bias, evaluation bias, and deployment bias [Suresh and Guttag, 2021] (See §3.2).

unsupervised learning the agent learns patterns in the input even though no explicit feedback is supplied [Russell and Norvig, 2002].

Publications during PhD

- **Michelle Seng Ah Lee**, Luciano Floridi, and Jatinder Singh. Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics*, pages 1–16, 2021: first author with subject matter guidance on ethical philosophy from Luciano Floridi and supervisory support [**Lee** et al., 2021]
- Heleen Jenssen, **Michelle Seng Ah Lee**, and Jatinder Singh. Practical fundamental rights impact assessments. *International Journal of Law and Information Technology*, 2022: second author contributing the perspective on AI and technology [Jenssen et al., 2022]
- **Michelle Seng Ah Lee** and Jatinder Singh. Risk identification questionnaire for unintended bias in machine learning development lifecycle. *Proceedings of the AI, Ethics, and Society Conference*, 2021b: first author with supervisory support [**Lee** and Singh, 2021b]
- **Michelle Seng Ah Lee**, Jennifer Cobbe, Heleen Janssen, and Jatinder Singh. Defining the scope of AI ADM system risk assessment. In *Research Handbook on EU Data Protection Law*. Edward Elgar Publishing, 2022: first author on a book chapter with subject matter input from Jennifer Cobbe and Heleen Janssen on referenced laws, regulations, and court cases and supervisory support from Jat Singh [**Lee** et al., 2022]
- **Michelle Seng Ah Lee** and Jatinder Singh. Spelling errors and non-standard language in peer-to-peer loan applications and the borrower's probability of default. In *In Proceedings of Credit Scoring and Credit Control Conference XVII*, 2021c: first author with supervisory support [**Lee** and Singh, 2021c]
- Emma Harvey, **Lee, Michelle Seng Ah**, and Jatinder Singh. **[Under review]** practical methods for measuring algorithmic fairness with proxy data. *AI and Ethics*, 2023: second author with contributions in study design, analysis, and insights [Harvey et al., 2023]
- Ari Ball-Burack, **Michelle Seng Ah Lee**, Jennifer Cobbe, and Jatinder Singh. Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In

Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 116–128, 2021: MSc supervisor to author, provided supervisory support with contributions in study design, analysis, and insights [Ball-Burack et al., 2021]

- **Michelle Seng Ah Lee** and Jat Singh. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021a: first author with supervisory support. Note: winner of Best Paper Award at ACM CHI [**Lee** and Singh, 2021a]

- **Michelle Seng Ah Lee**, Luciano Floridi, and Jatinder Singh. Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics*, pages 1–16, 2021: first author with subject matter guidance on ethical philosophy from Luciano Floridi and supervisory support from Jat Singh [**Lee** et al., 2021]

- Wesley Hanwen Deng, Manish Nagireddy, **Lee, Michelle Seng Ah**, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. Exploring how machine learning practitioners (try to) use fairness toolkits. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022: a follow-up study from my Landscape and Gaps paper in which I provided guidance, analysis, and insights [Deng et al., 2022]

- Jennifer Cobbe, **Michelle Seng Ah Lee**, and Jatinder Singh. Reviewable automated decision-making: A framework for accountable algorithmic systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 598–609, 2021: contributed AI and technology specific guidance on what factors lead to reviewability [Cobbe et al., 2021]

- Kornel Lewicki, **Lee, Michelle Seng Ah**, Jennifer Cobbe, and Jat Singh. [**Under review**] out of context: Investigating the fairness concerns of "artificial intelligence as a service". In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023: MSc supervisor to Kornel Lewicki, contributed insights and study design [Lewicki et al., 2023]

- Seyyed Ahmad Javadi, Richard Cloete, Jennifer Cobbe, **Michelle Seng Ah Lee**, and Jatinder Singh. Monitoring misuse for accountable 'artificial intelligence as a service'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 300–306, 2020: contributed AI and technology specific guidance on how AI-as-a-service monitoring would work in practice [Javadi et al., 2020]

- **Michelle Seng Ah Lee**, David Watson, , and Zhe Feng. [IN PROGRESS, PENDING SUBMISSION] Fairness under uncertainty in sequential decisions. In *Proceeding of the 39th IEEE International Conference on Data Engineering (ICDE)*, 2023: first author responsible for the idea origination, framework design, and initial visualisations. I will be looking to co-authors for support in the full python package implementation, which is ongoing at the time of submission [**Lee** et al., 2023]

Chapter 1

Introduction

Machine learning (ML) algorithms are increasingly used to inform high-impact decision-making, from credit risk evaluation to hiring to criminal justice. Due to their usage of large, non-traditional data sources, frequent re-training cycles, and complex methods, it has become difficult to detect when ML design may be misaligned to the designer's intent, an organisation's legal obligations, and societal expectations. One particular concern has been the presence of unintended biases and unfair discrimination, such as against certain racial and gender groups [Hutchinson and Mitchell, 2019; Mehrabi et al., 2021].

There has been growing consumer awareness of ethical considerations of ML, as the perceived "unfairness" of algorithm-assisted decisions are becoming headline news. A model used to assign a recidivism score in the criminal justice system sparked controversy when it was accused of overestimating the risk of black defendants [Feller et al., 2016]. Applecard by Goldman Sachs was investigated by a regulator after customer complaints of women receiving lower credit limits than men with the same credit standing [Vigdor, 2019]. Recruiting tools at a technology company were reportedly biased against women [Dastin, 2018].

In managing the reputational risk of using algorithms, organisations should also be wary of appearing superficial or tokenistic in their approaches. The increasing public focus on the risks and harms of unfair bias in algorithmic decision-making has led to fairness becoming a lightning rod for scrutiny and criticism of usage of these technologies more broadly, and criticism has been levelled at various organisations for "Fair-washing": selecting fairness metrics to promote the false perception that a model respects some ethical values [Aïvodji et al., 2019].

Fairness issues in ML can also result in regulatory and legal risks for organisations. With increasing public scrutiny, regulators have started actively issuing guidance documents outlining their expectations, as well as developing and implementing new regulation [Information Commissioner's Office, 2017; Government of the Netherlands]. The European Union Artificial Intelligence Act proposed in April 2021 is still in drafting stage; however,

in its current format, it seeks to set out coordinated rules in Europe for development and usage of AI systems in the market [Veale and Borgesius, 2021]. The international human rights legal framework, codified in the Universal Declaration of Human Rights and supported by other treaties and documents, establishes the principles of non-discrimination on the basis of certain features such as sex, race, language, or religion [Assembly et al., 1948]. From a legal standpoint, the approach in automating "fairness testing" appears incompatible with the requirements of EU non-discrimination law, which relies heavily on the context-sensitive, intuitive, and ambiguous evidence [Wachter et al., 2021]. Although Wachter et al. propose a new metric to counteract this, the metric assumes that what type of equality should be achieved is known and agreed, with limited guidance on what is fair in each context.

Given these risks, organisations have an incentive to ensure their ML systems are aligned to their values and ethical principles; however, there is limited consensus in academic literature on how well-intentioned organisations can avoid unintended and unfair biases throughout their ML pipelines. Holistic governance of fairness challenges in ML systems requires an understanding of the legal, organisational, and technical context of their deployment. In this thesis, we provide approaches, tools, and methods to tackle fairness considerations in an end-to-end ML development lifecycle – not as a one-size-fits-all solution, but rather, as starting points that can be tailored to each context and use case.

1.1 Definitions of AI, ML, and related concepts

For the purpose of this thesis, we first define **machine learning (ML)**, **artificial intelligence (AI)**, and **automated decision-making (ADM)**, also found in the Glossary. Stuart Russell and Peter Norvig, in their highly-prominent textbook *Artificial Intelligence: A Modern Approach*, define AI as "the designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment" [Russell and Norvig, 2002]. Percepts and actions are ultimately data flows, and the algorithm learns patterns from data, in contrast to systems in which a human designer explicitly hard-codes the 'rules'. **Machine learning (ML)** refers to a set of techniques used in AI to detect and extrapolate patterns from data and includes **supervised learning**, **unsupervised learning**, **semi-supervised learning**, and **reinforcement learning** [Russell and Norvig, 2002]. Because these algorithms often process data for the purpose of deriving insights for decision-making, AI is linked to Automated Decision-Making (ADM) as one of the techniques that may be used. ADM is the automation of business processes, while AI is the technique used in the system. ADM, such as credit scoring systems, often has requirements to involve a human reviewer, a common safeguard found in various regulatory instruments and provisions [Binns, 2019]. ADM is directly

referenced in the EU General Data Protection Regulation (GDPR) [European Union]. The relationship between these definitions is further explored in §3.1.

The scope of what constitutes AI is unclear in the above definitions, particularly when such definitions are being considered in the context of organisational risk management. Some argue that AI methods encompass relatively simple algorithms, such as logistic regression or rules-based models, but others limit the definition of AI to more complex models, such as deep neural networks [Kaplan and Haenlein, 2019]. This debate can be summarised in a frequently misquoted "AI is whatever has not been done yet," which the original author corrected to be "intelligence is whatever machines haven't done yet" [Haigh, 2011]. As such, the definition of AI varies without a clear consensus.

For the purpose of this thesis, we indicatively define the key terms as follows:

- **model**: a formal, usually quantitative, representation of a real-world phenomenon by which a prediction, decision, or recommended action is derived, given known factors and assumptions. A model can be based on data and/or expert knowledge, by humans and/or by automated tools like machine learning algorithms;
- **algorithm**: computational method, formula, or procedure;
- **technique**: method used in the technical design, build, and testing of the algorithm;
- **process**: a series of logical / ordered operations involved in decision-making, encompassing both the technical and business actions taken;
- **data processing**: any operation or set of operations which is performed on personal data;
- **system**: a set of interacting data, algorithm(s), and/or model(s) to form a technical workflow or product, e.g. a facial recognition algorithm that triggers an identity verification model; and
- **profiling**: automated personal data processing with the objective of evaluating personal aspects about a natural person, including to analyse or predict aspects of their performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements (GDPR [European Union], Article 4(4) [European Union], Recital 71 [European Commission, c]). A system may perform profiling of people as a part of ADM and /or using AI. We discuss the interplay between AI and ADM in Chapter 3.

It is difficult to achieve consensus on some of these terminologies, especially AI – there are different opinions on each term's scope and delineation. We disentangle these definitions further in Chapter 3, highlighting the limitations of using specific terminology

where considerations are broadly relevant. All definitions for the purpose of this thesis are listed in the Glossary.

1.2 Key gaps in academic literature and practical challenges

In response to both public and regulatory scrutiny against AI systems, especially those involved in ADM, there has been a proliferation of computer science literature on algorithmic fairness aiming to quantify the deviation of their predictions from a formalised metric of equality between groups (such as male and female). Dozens of metrics of fairness have been proposed, prompting efforts to disentangle their differences and rationale [Verma and Rubin, 2018]. However, there is little consensus on how these challenges should be addressed.

The three gaps that motivate this thesis are:

1. the gap between the narrow algorithmic definitions and the inherent complexity, context-specificity, and subjectivity of "fairness";
2. the gap between the existing tools and the practitioners' requirements for end-to-end, context-specific solutions; and
3. the gap between the unambiguous fairness tests and real-world uncertainties.

The first gap is how fairness is defined narrowly in fair ML literature, whereas it is a complex concept debated for millennia across disciplines. The second gap exists because the toolkits built to implement these definitions for practical usage are not aligned to the practitioners' requirements. Their demand for an end-to-end solution that can be adapted to their own use cases is yet unmet. Finally, the fairness tests give seemingly unambiguous answers to whether or not the system is fair without consideration of the real-world uncertainties that plague each stage of the ML development lifecycle.

1.2.1 Gap between the narrow algorithmic definitions and the inherent complexity, context-specificity, and subjectivity of "fairness"

The scope of "fairness" defined in academic literature is often extremely narrow in order to "fix" it using other algorithmic approaches. Such a reductionist approach is at odds with the inherent complexity of what it means to be fair, a topic of debate across disciplines such as ethical philosophy and welfare economics. Many fairness definitions that are formalised in academic literature are mathematically incompatible, which is reflective of

the fact that the philosophical perspectives from which they are derived are often morally irreconcilable (shown in §2.3.2). Therefore, one fairness definition cannot fit all contexts.

While "fairness" may be a concept we feel we understand intuitively in a humanistic sense, it is a challenging one on which to gain consensus and even more complex to operationalise in practice. Fairness cannot be reduced to a box-ticking exercise in a checklist, because whether or not an algorithm is fair depends heavily on the context and perceived benefit or detriment of outcomes. What is fair based on one metric, or in one jurisdiction, from one cultural and social perspective, or in the opinion of one stakeholder, may be considered unfair in another context. Difficult choices must sometimes be made, and decision-makers must decide whether they are satisfied with achieving legal compliance (for example, with anti-discrimination obligations), or whether they feel a greater responsibility around fairness related to concepts such as equity and social justice.

This disconnect between real-world complexity of what it means to be fair and the proposed axiomatic fairness definitions is not new. Hutchinson and Mitchell (2019) warn of the gap between the unambiguous formalisation of fairness metrics and the contextual and practical needs of society, politics, and law. They compared the recent surge in ML fairness research to literature from the 60s and 70s, which fizzled with the following conclusion:

> "no statistic that could unambiguously indicate whether or not an item is fair was identified. There were no broad technical solutions to the issues involved in fairness" [Cole, 1973].

The human yearning to measure and quantify the world around them is enshrined across history, beyond computer science [Vincent, 2022; Abebe et al., 2020]. While the methodology of measurement has its limitations, being able to quantify them means that they can be tracked and discussed. Unlike unfairness in human decision-making that may be mired in cognitive biases, algorithmic decision-making can be assessed for unfairness and debated on its design. We acknowledge the importance of quantification, but it is sensitive to various contextual dimensions.. We discuss how to understand the holistic ethical considerations beyond fairness in Chatper 2.

Fairness issues should be addressed – not only at an algorithmic level – but also at the *system-level* including the process and the people. This is enshrined in **General Data Protection Regulation (GDPR)**: personal data should be processed securely by means of "appropriate technical and organisational measures," including *risk analysis* and policy changes [European Union]. Fairness and potential discrimination are potential legal risk factors. Tackling the contextual considerations of a risk is aligned to the regulatory expectations of GDPR compared to only addressing the technical procedure. Existing tools to address fairness issues, however, have been critiqued for falling short of practical requirements. We address this gap in this thesis by deriving lessons from other disciplines,

such as from ethical philosophy and welfare economics, to approach fairness – not as a solely algorithmic testing exercise – but as a holistic ethical review of how a model is successfully or unsuccessfully accomplishing the use case's competing objectives.

1.2.2 Gap between the existing tools and the practitioners' requirements for end-to-end, context-specific solutions

The "toy" case studies, data sets, and tools developed used for fairness testing have limited similarity to real-life complexities. In recent years, a number of **fairness toolkits** have been introduced, providing the means for testing the algorithm's predictions against various fairness definitions. The open source fairness toolkit landscape so far reflects the reductionist understanding of fairness as mathematical conditions, as the implementations rely on narrowly defined fairness metrics to provide "pass/fail" reports. These toolkits can sometimes give practitioners conflicting information about an algorithm's fairness, which is unsurprising given that it is mathematically impossible to meet some of the fairness conditions simultaneously [Kleinberg et al., 2016]. This is reflective of the conflicting visions of fairness espoused by each mathematical definition and the underlying ethical assumptions [Binns, 2020].

Our recent paper surveying the fairness toolkit landscape [**Lee** and Singh, 2021a] found there were significant gaps between ML practitioner needs and the toolkits' features. Other studies involving ML practitioners have similarly identified the unmet demand for domain-specific and contextual factors to be closely considered to improve algorithmic fairness [Veale et al., 2018]. In many domains, practitioners claim that fairness cannot be understood in terms of well-defined quantitative metrics [Holstein et al., 2019].

The tools have limited coverage of the end-to-end lifecycle. In our study of practitioners' views of fairness toolkits [**Lee** and Singh, 2021a], the interviewees emphasised the apparent focus of the toolkits on the model building and evaluation process as compared with the remaining model lifecycle. One ML engineer stated, "Each section of the model building pipeline is important – testing your training data, representation, model output, proxy variables, etc... no tool has an end-to-end 'this is what is going on in your system.' " This echoes previous findings [Holstein et al., 2019] that the scope of current tools are limited in their coverage of the ML pipeline.

Fairness toolkits aim to be widely accessible, drawing attention to common fairness considerations, and encouraging and supporting practitioners to consider, assess (and therefore mitigate) their algorithms in leading to unfair outcomes. However, without a consideration of the *relevant context* in the socio-technical system surrounding the algorithm, these tools risk engendering false confidence in flawed algorithms. Different considerations come into play for each use case. That is, organisations should not rely solely

on one-dimensional algorithmic fairness metrics to account for its ethical concerns. These narrow applications of fairness could mislead organisational strategy, risk management, and policies. We address this gap in this thesis through an end-to-end consideration of the full lifecycle (design, build, test, and monitor) with tools and approaches that reveal context-specific fairness issues.

1.2.3 Gap between the unambiguous fairness tests and real-world uncertainties

The scarcity of practical solutions proposed for fair ML is exacerbated by the limited consideration for the system-level considerations, especially the layers of uncertainties. In our study of fairness toolkits, one industry practitioner commented that the fairness test results *"make everything look clear-cut, which it really isn't 'in the wild' "* [**Lee** and Singh, 2021a]. There are layers of uncertainty in each stage of an ML development lifecycle that have implications for the system's fairness, which are overlooked in the fairness tests.

In addition, most fairness metrics are defined in a **supervised learning** setting [Verma and Rubin, 2018]. These are incompatible with how decisions are made in many real-life high-stakes settings that require dynamic decisions, rather than static predictions. In domains such as insurance pricing, fraud detection, hiring, and lending, predictions are not evaluated in a one-off batch processing of data; rather, a decision is made for each individual or a batch of individuals, and the outcome of that decision informs future policies. Each decision in such a setting is made under uncertainty, which is overlooked in supervised ML. For example, a bank cannot know whether a denied loan would have been repaid, and it may have less data about previously marginalised and financially excluded populations. We address this gap in this thesis through a consideration of uncertainties in ML systems. Uncertainty in sequential decision-making through ML should be taken seriously and actively considered in understanding system-level fairness.

1.3 Our contribution

The objective of this thesis is to address these key gaps that hinder fair ML development in practice: the gap between the narrow problem-solving in past literature to "fix" unfairness and the need for a more holistic approach, the gap between proposed tools and their contextual, end-to-end applicability, and the gap between the unambiguous formalisation of fairness in academia and the inherent uncertainties in real-world decision-making. To this end, we present an end-to-end methodological guidance throughout the lifecycle of fair ML, which takes into account the contextual nuances, practical considerations, and a holistic understanding of ML ethics.

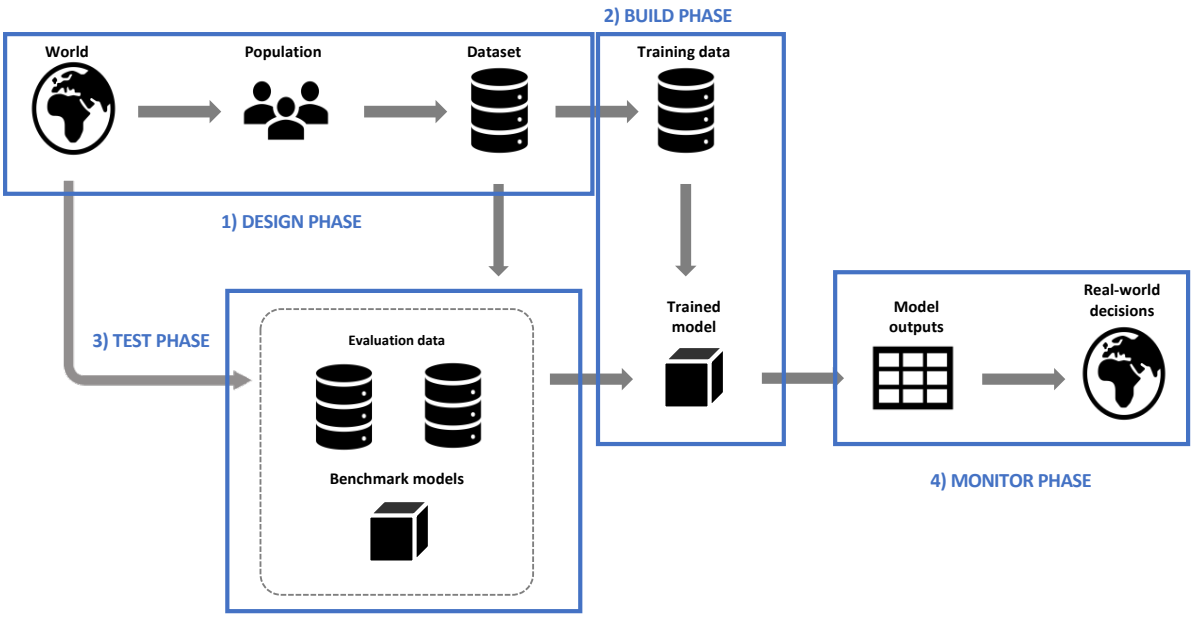


Figure 1.1: Phases of ML development lifecycle

The thesis structure will be aligned to the phases of an ML development lifecycle, shown in Figure 1.1. Much of this thesis is aimed at the ML developer, who would be the target user of the approaches and tools we propose, such as the bias identification questionnaire. However, we stress throughout our thesis the importance of an understanding of the complexity of defining fairness and the shared responsibility across each organisation on ensuring every ML system built is aligned to its ethical values. In this thesis, we focus on a subset of stakeholders that make decisions throughout an ML development lifecycle. This includes not only the technical stakeholders, such as the ML developers, but also the business leaders, business risk functions, policy-makers, and regulators who may set expectations and limitations on how ML should be designed and developed. Thus, we aim to keep our technical language accessible such that an audience beyond the developers and the fair ML community would find this thesis a useful reference point. We now outline the phases of ML development lifecycle, which form the structure of this thesis.

1. **Design** In the design phase, the developer identifies a data set that represents a sample of the global population and decides on a model to build. For example, a bank may have data on its past borrowers and aim to build an algorithm to predict loan outcomes for future borrowers. It is important to understand from the onset whether there are ethical risks, including potential for unfair outcomes, and determine whether to proceed with the model build based on the assessment of those risks against the potential benefits. This chapter will show the limitations in computer science literature in assisting this assessment and propose a new approach: "Key Ethics Indicators" (KEIs).

2. **Build** In the build phase, the developer compiles a training data set for the model. This includes feature selection, feature engineering, and model selection, all of which involves decisions that may introduce unintended and potentially unfair biases into the ML model. This chapter will propose a "bias identification questionnaire" as a resource for developers in order to design targeted mitigation strategies. We present the results validating the questionnaire's effectiveness through surveys of industry practitioners.

3. **Test** Scholars have built toolkits to facilitate testing for fairness. This chapter covers the landscape of open source fairness toolkits and identifies the major gaps that hinder their adoption in practice through a mixed-methods study of industry practitioners. We end with a case study of how KEIs can provide a more fit-for-purpose evaluation of an ML model.

4. **Monitor** In the monitoring phase, the developer has deployed the ML model, which is being re-trained in a live environment. There is limited prior work on ensuring fairness under the uncertainty introduced into the ML model in sequential decision-making process. For example, the lender does not know whether denied loans would have defaulted or would have been repaid. In a domain area with a history of discrimination, there would be greater uncertainty around previously marginalised and excluded groups. This chapter proposes a taxonomy of six types of uncertainty in an ML pipeline and a new approach that takes uncertainty into account when ensuring fairness in sequential decision-making.

The following chapters will cover the entire ML lifecycle, presenting the key gaps in academic literature and proposing practical methods and tools that are validated through our studies. By bringing together literature from ethical philosophy, welfare economics, and computer science, Chapter 2 identifies the disparity in how fairness is defined in each domain area (the first gap in §1.2.1) and proposes an approach called Key Ethics Indicators ("KEIs") that combines lessons from each discipline to holistically assess the ethics of a proposed ML system. Chapter 3 and Chapter 4 addresses the second gap in §1.2.2 by supplementing our KEI approach with a questionnaire designed to identify biases in the end-to-end ML development lifecycle. Finally, Chapter 5 proposes a taxonomy of uncertainty across the lifecycle and ways in which uncertainty can be accounted for in sequential decision-making using ML predictions (third gap in §1.2.3). Our hope is that this thesis will contribute to a foundation for end-to-end guidance, analysis, and new proposals for industry practitioners, regulators, and academics on how fairness can be considered and operationalised in practice.

Chapter 2

ML Design: contextually defining algorithmic fairness

Introduction

In the design phase, it is important to ask 1) whether there are any fairness issues at hand and 2) whether to proceed with the model build or not based on a risk-benefit analysis. Some ML systems may not have apparent or directly consequential fairness considerations, such as financial forecasting models that have no human component or interface. However, many ML systems that inform decision-making, from credit risk evaluation to insurance pricing to hiring to criminal justice, do raise significant fairness issues. Even seemingly non-personal data can encode information about individuals that – when used to build models – may result in unfair outcomes: for example, one's handwriting data can divulge one's nationality [Nag et al., 2018]. The developer and relevant organisational stakeholders need to determine whether the potential benefits of a proposed ML system outweigh the potential risks. A landmark example of when the risks outweighed the benefits on a societal level was the moratorium on facial recognition in San Francisco by police and other government agencies, where it was determined that the risks to privacy outweigh the potential safety benefits [Conger et al., 2019]. Despite the clear need to discuss whether a system is worth building at all before commencing development, there is limited guidance in academic work in assisting this go/no-go decision in design phase, as it primarily concerns itself with testing models that have already been built.

This chapter sets the scene on how algorithmic fairness is defined in existing literature, contesting the key gaps in how quantification of fairness can simplify the notion of fairness. Defining fairness in a manner that befits its contextual nuances is the key to understanding why suggestions from the algorithmic fairness literature have not been widely adopted in practice. In order to understand the system – not only as a technical algorithm – but also as a conscious design on how to model a real-world social phenomenon, it is

important to look beyond computer science literature. Drawing from welfare economics and ethical philosophy, we will show the axiomatic fairness definitions are inconsistent with the necessarily subjective and holistic notion of fairness in other disciplines. We end the chapter by proposing a new approach of using "Key Ethics Indicators" (KEI) in the design phase to determine whether the potential benefits outweigh the risks of building the ML system. In future chapters, we will return to the Key Ethics Indicator approach to demonstrate how defining KEIs in the design phase helps inform decisions throughout the build, test, and monitoring phases. Designing a fair system implies not only defining what type of model we want to build but also what type of world we want to see reflected in our model.

2.1 Defining key terms

Before we discuss the limitations of fairness metrics in computer science literature, it is important to first define key terms for the purpose of this thesis. This section will examine notions that are relevant to our discussions. While these terms do not comprehensively cover all relevant aspects of algorithmic ethics, they clearly demonstrate the limitations of mathematical fairness formalisations in capturing necessary information about the algorithmic system. These definitions are also included in the Glossary.

2.1.1 Ethics in AI

Ethics as it is used in AI and algorithmic settings may be described as the systematic conceptualisation of 'right' and 'wrong' behaviour, which are often reflected in an accepted set of rules and principles. Over 160 guidelines related to data and AI ethics have been proposed globally, while various organisations have selected combinations of principles into the multitude of "AI ethics" frameworks [AlgorithmWatch, 2019]. Five common themes have been identified across these sets of principles: beneficence, non-maleficence, autonomy, justice, and explicability, defined below in Table 2.1 [Floridi et al., 2018; Floridi and Cowls, 2022].

While the focus of this thesis is *justice* and the relevant notion of *fairness*, we will show that justice and fairness must be contextualised alongside other ethical principles, as there may be trade-offs among them. In addition, while there have been a multitude of principles proposed, scholars have argued that they are unhelpful without *operationalisation* into practice [Canca, 2020; Lee et al., 2020].

In this chapter, we demonstrate how this obstacle can be overcome by operationalising the principles into Key Ethics Indicators. In other words, the principles of fairness and beneficence need to be translated into real-world implications to help inform the decision on which model is best aligned to the organisation's values. On their own, these sets of

Common AI ethics principle	Definition in Floridi et al. (2018)
Beneficence	Promoting well-being, preserving dignity, and sustaining the planet
Non-Maleficence	Privacy, security and "caution" around the "upper limits on future AI capabilities"
Autonomy	The power to decide (whether to decide)
Justice	Promoting prosperity and preserving solidarity
Explicability	Enabling the other principles through intelligibility and accountability

Table 2.1: Five common themes in AI ethics framework cited in Floridi et al. (2018)

principles, in whatever permutation they are presented, do little to guide the trade-offs between competing principles in a use case.

2.1.2 Justice, equality, and equity

A study of proposed ethical principles finds that the different countries' and organisations' understanding of **justice** varies for each document, from the elimination of discrimination to promoting diversity to shared prosperity [Floridi and Cowls, 2022]. Justice is defined in this thesis in accordance with legal and organisational science literature, with justice denoting adherence to the standards agreed upon in society (for example, based on laws) and fairness as a related principle of an evaluative judgement of whether a decision is morally right [Goldman and Cropanzano, 2015].

In line with this definition, fairness is inherently subjective. The concept is based on the notion of **equality**: the egalitarian foundation that humans are fundamentally *equal* and should be treated *equally*. Fairness in a sociological sense defines the criteria under which some people "deserve" a limited resource more than others. However, how equality should be measured and to what extent it is desirable have been a source of debate in both philosophical ethics from a moral standpoint (§2.3.2), and welfare economics from a market efficiency standpoint (§2.3). In fact, a study of human behaviour showed people prefer fairness over equality when they are at odds with one another [Starmans et al., 2017]. The criteria for distribution of limited resources are inherently subjective and depend heavily on the ethical values of the decision-maker, the surrounding circumstance, and context. For example, Aristotle wrote that if there are fewer flutes available than people who want to play them, it is fair that they should be given to the best performers [Aristotle and Sinclair, 1962]. Various possibilities for these criteria from ethical philosophy literature will be discussed further in §2.2.2.

The core aim of **equity** on the other hand, is to reduce avoidable inequalities, such as the absence of systematic disparities in health between social groups who have different levels of underlying social advantage/disadvantage [Braveman and Gruskin, 2003]. In other

words, the poor should be just as healthy as the rich and privileged. Driving equitable outcomes involves targeting support at the marginalised, disadvantaged, and vulnerable communities, even if this results in these groups receiving more or better support than what someone else gets in a system. Some philosophical perspectives on fairness focus on equity (discussed in§2.3.2). There is general consensus among economists that reducing extreme inequalities is beneficial for all [Johansson, 1991].

2.1.3 Discrimination and protected characteristics

Fairness should be distinguished from *discrimination.* From a legal standpoint, **discrimination** refers to the notion that *protected characteristics* should not result in a relative disadvantage of deprivation. Protected characteristics are those commonly referenced and reflected in non-discrimination laws, such as race and ethnicity, gender, religion, age, disability, and sexual orientation, given these personal demographic features are central to discussions on algorithmic fairness. Non-discrimination laws aim to not only prevent ongoing discrimination but also to change societal policies and practices to achieve more substantive equality – an aim which is described as incompatible with some fairness metrics [Wachter et al., 2020]. While legal analysis is outside the scope of this thesis, we refer to protected characteristics as those commonly referenced and reflected in non-discrimination laws, such as race and ethnicity, gender, religion, age, disability, and sexual orientation, given these personal demographic features are central to discussions in the algorithmic fairness literature. We also refer to **direct discrimination**, which concerns differential treatment based on a protected characteristic and **indirect discrimination**, where a rule of policy applying to all produces a negative impact on a protected group [Wachter et al., 2020].

2.1.4 Bias

Bias is a colloquial reference to prejudice against one person or group, especially in a way which could be considered to be unfair [**Lee** et al., 2021]. This is often a confusing term in computer science, as "bias" in machine learning, statistics, and econometrics refers to a type of error in a learning algorithm that results in under-fitting the data (compared to *variance*), which results from using too simple of a model to represent a complex relationship [Hastie et al., 2009; Mayson, 2018]. It is used in cognitive sciences as a systematic error in thinking when people are processing information, which are evolutionary functions that form "shortcuts" for the human mind [Haselton et al., 2015]. In clinical research, "statistical bias" refers to the systematic difference between results and facts that is introduced in the design or conduction of research [Tripepi et al., 2008].

For the purpose of this thesis, we define bias as unintended and potentially harmful

skewing of algorithmic predictions. Suresh and Guttag (2019) created an algorithmic bias taxonomy focusing on those that cause "unintended consequences," categorising all biases into six distinct categories spanning the AI development pipeline: historical bias, representation bias, measurement bias, aggregation bias, evaluation bias, and deployment bias [Suresh and Guttag, 2021]. This will be further discussed in Chapter 3 (ML Build).

The above concepts of ethics, justice, discrimination, and bias are relevant to the definition of fairness but are often used inter-changeably, such as bias to denote unfairness [**Lee** and Singh, 2021b]. It is important for the academic community to be precise in our language because these terms are over-loaded due to usage across disciplines.

2.2 Fairness: definitions in computer science

Fairness is one such over-loaded term with a long history of discussion across legal, ethical, philosophical, and economic literature. In computer science, Scholars have often formulated fairness in a quantitative way and attempted to maximise it. In the context of this thesis, we refer to *algorithmic fairness* specifically, as the term is also used in other settings in computer science, such as for resource allocation in schedulers for operating systems or networks [Kumar and Kleinberg, 2000]. For example, *maximin fairness* in the distribution of computer network's bandwidth denotes the prioritisation of smaller flows [Denda et al., 2000], which is interestingly derived from Rawlsian equality of opportunity and the maximin rule (See definition of equality of opportunity in both ethical philosophy and algorithmic fairness literature in §2.3.2). Another fairness definition unique to this domain is *proportional fairness*, which balances between maximising total throughput of the network and at the same time allowing all users at least a minimal level of service [Kelly, 1997]. These are not in scope for our thesis. In addition, while we refer to some legal scholarship, a discussion of the evolution of non-discrimination laws and their philosophies is out of scope for this thesis. Future work may consider the case law related to fairness and discrimination to reveal further insights into how they are defined in various domain areas.

A quantitative approach to algorithmic fairness has its benefits: to enable it to be measured and improved upon using computational techniques. However, existing mathematical definitions of fairness in computer science literature should be calculated while keeping in mind the nuances and context-specificity present in philosophical discourse. While they are loosely derived from a notion of egalitarianism, the definitions have little resemblance to the complex philosophical notions they aim to represent. We link our contribution to other work connecting philosophical discourse to computer science literature on fairness [Binns, 2020; Heidari et al., 2019] and expand on them. In drawing links between political philosophy and the fair ML literature, Fazelpour and Lipton 2020 argue

that these metrics that aim for an "ideal world" in which parity exists fail to "account for the mechanisms by which our non-ideal world arose," which can lead to "misguided policies." Our work provides a broad overview of different ethical paradigms of what is ideal, demonstrating not only how they conflict with one another in theory but also what each view of fairness prioritises in practice.

In this section, we first iterate through the key definitions of fairness in computer science literature and connect them to real-life intuitions and implications. We then show their morally irreconcilable assumptions on the desirability of equality by connecting the definitions to relevant philosophical doctrines and demonstrating the gaps. While there may be other metrics that represent tweaks to those presented here, these represent the foundational metrics that equalise the outcome and prediction and include all those mentioned in literature attempting to give a landscape of the metrics (e.g. [Verma and Rubin, 2018]). The different names given to the same metric is also discussed previously [Verma and Rubin, 2018] to be potentially confusing, exacerbating the challenge of selecting a fairness metric.

2.2.1 Use case: Defining fairness in mortgage lending

To bring the fairness definitions to life, we will walk through a use case: a lender building a model to predict a prospective borrower's risk of default on a loan, while attempting to assess whether the model is fair across black and white applicants. This is one of the domain areas with a known history of racial discrimination; in the U.S., mortgage lenders have long been accused of illegally and unfairly denying loans to black applicants [Fuster et al., 2022].

In this case, the False Positives (FP) represent lost opportunity (predicted default, but would have repaid), and the False Negatives (FN) represent lost revenue (predicted repayment, but defaulted). The calculations of error rates used in the metrics are defined below, with some of the most commonly cited fairness definitions in Table 2.2 and Table 2.3:

- True Positive Rate (TPR) = TP/(TP + FN)
- True Negative Rate (TNR) = TN/(FP + TN)
- False Positive Rate (FPR) = FP/(FP + TN) = 1 – TNR
- False Negative Rate (FNR) = FN/(FN + TP) = 1 – TPR
- Positive Predictive Value (PPV) = TP/(TP+FP)

Note that among the numerous definitions in Table 2.2 and Table 2.3, there are difficulties in deciding which metric is most appropriate for each use case [**Lee** and Floridi, 2020]. Is a 3% increase in positive predictive parity preferable over a 5% increase in equal

Fairness metric	Equalising	Intuition (Example)	Practical Objective
Maximise total accuracy	N/A	The most accurate model gives people the loan and interest rate they 'deserve' by minimising errors	**Minimise unaffordable loans**: Defaults are harmful to both the business and the borrower; **Allocative efficiency**: limited capital goes to those most likely to repay; **Portfolio risk / financial inclusion**: For a lender with limited capital, reducing portfolio risk of defaults enables the lender to give more loans to more people for those who were at the borderline between approval and denial; **Reduced missed revenue opportunities (reduce overall False Negatives)**: Denying loans that would have been repaid represents missed revenues; **Minimize financial loss due to default (reduce overall False Positives)**: approved loans that default is expensive
Demographic parity, group fairness, disparate impact [Feldman et al., 2015]	Outcome	Black and white applicants have the same loan approval rates	**Improve racial equity**: if we assume a level playing field, i.e., black and white applicants have the same default risk distribution, this metric is sensible. Note that if there are legitimate differences in default risk: for example, in income, then engineering group fairness will reduce the accuracy of the model. These legitimate differences may be linked back to underlying social and racial inequalities and injustices.
Equal opportunity / false negative error rate balance [Hardt et al., 2016]	FNR	Among applicants who would default, both black and white applicants should have similar rate of their loans being denied	**Reduce unfair loan denials for black applicants**: If creditworthy black applicants have higher rates of loan denials due to model's errors, this may be seen as unfair. Note that False Negatives may be impossible to measure in this example because the lender would not have information on whether a denied loan would have been repaid or would have defaulted. This requires assumptions or estimations.
False positive error rate balance / predictive equality [Chouldechova, 2017]	FPR	Among applicants who are credit-worthy and would have repaid their loans, both black and white applicants should have similar rate of their loans being approved	**Minimise harm to black borrowers**: if black applicants are more often incorrectly predicted to repay but eventually default, then the inaccuracies are disproportionately harming black applicants who are left with unaffordable loans.

Table 2.2: Fairness metrics and their intuitions (Part 1/2) - what each metric equalises with a lending example, translated into practical objectives for the lender

Fairness metric	Equalising	Intuition (Example)	Practical Objective
Equal odds [Hardt et al., 2016]	TPR, TNR	Meets both of above conditions	**Minimise racial differences in errors**: This metric aims to minimize all differences in errors, aggregating the metrics from false positive and false negative error rate balances. It is important to note that equal opportunity, predictive equality, and equal odds are known as "bias-preserving," assuming that any underlying inequalities between groups are benign. Where there are structural discriminatory biases embedded in the data, these metrics would not address them. See [Wachter et al., 2020] for a discussion on how this relates to US and EU discrimination law.
Positive predictive parity [Chouldechova, 2017]	PPV	Among credit-worthy applicants, the probability of predicting repayment is the same regardless of race	**Ensure equal probability of correct predictions between black and white applicants**: This is similar to FNR but looks at probability rather than a binary outcome. Evaluating based on probability, rather than binary outcome, differentiates between those who were a borderline case from those who were a clear approval or rejection.
Positive class balance [Kleinberg et al., 2016]	Average probability of positive class	Both credit-worthy white and black applicants who repay their loans have an equal average probability score	**Avoid over-estimating the creditworthiness of white applicants**: if the probability of white applicants' approval is higher than probability of black applicants' approval among those who repaid their loans, it could indicate white applicants' creditworthiness is inflated
Negative class balance [Kleinberg et al., 2016]	Average probability of negative class	Both white and black defaulters have an equal average probability score	**Avoid over-estimating default risk of black applicants**: By contrast, this metric would indicate that black applicants' default risks are comparatively inflated, and the true risk may be lower
Counterfactual fairness [Kusner et al., 2017]	Prediction in a counterfactual scenario in which the person had a different protected feature	For each individual, if he/she were a different race, the prediction would be the same	**Demonstrate "race-blindness"**: This metric could show an individual that if he/she were a different race, the outcome would have been the same. However, this is difficult to show in practice. For example, if an applicant were white, his/her income, neighbourhood, occupation, and financial history may also be different. This issue of proxies is further discussed in Chapter 3. Note: **proxies** are features that encode information about sensitive features (See §3.6).
Individual fairness [Dwork et al., 2012]	Outcome for 'similar' individuals	For each individual, he/she has the same outcome as another 'similar' individual of a different race	**Explain why a loan was approved or denied**: An organisation may argue that the decision was fair because other "similar" applicants of another race had the same outcome. However, it is difficult to define "similarity" that is independent of race.

Table 2.3: Fairness metrics and their intuitions (Part 2/2) - what each metric equalises with a lending example, translated into practical objectives for the lender

odds? Moreover that many of these metrics cannot be satisfied at the same time [Kleinberg et al., 2016], it is not intuitive on which metric best represents the lender's interests. From a legal standpoint, while these metrics may provide a reason to suspect unfairness exists, a lack of quantitative parity does not in itself constitute discrimination [Hellman, 2020].

The tensions between these approaches both from a mathematical standpoint and from a moral and value standpoint can only be resolved if the metric is explicitly linked to the contextual objectives. What trade-off between our competing objectives is the decision-maker comfortable with and capable of justifying to all associated stakeholders (e.g. regulators)? Our KEI approach in §2.5 provides a process of contestation, in which these tensions are revealed and debated. Rather than selecting a generic metric that is a poor abstraction of philosophical ethics, KEIs can present the decision-makers with a concrete "menu" of options. There may not be one optimal solution, but this process makes the value judgements explicit.

These issues will be further discussed in §2.3, where we will link each fairness metric to its philosophical origin and address the gaps. The gaps, in particular, demonstrate the important nuances that cannot be captured in fairness metrics that must be considered in the model development process. In §2.2.2, we challenge the types of inequalities that the fairness metrics assume are acceptable vs. unacceptable.

2.2.2 Flawed assumption: simplicity of separating acceptable inequalities from the unacceptable

Fairness metrics assume it is possible to separate acceptable inequalities from the unacceptable. We challenge this assumption by discussing the complexity of the debates on equality in ethical philosophy. Note that these metrics are aimed at a class of machine learning algorithms that are supervised, i.e. with a known outcome, and for classification purposes, i.e. for a discrete outcome (e.g. default vs. repayment) rather than a continuous outcome (e.g. amount repaid). These algorithms aim to identify the features that are associated with the outcome of interest. For example, a loan applicant with higher income is more likely to be approved due to income's association with higher ability to repay. In this case, differences in socioeconomic status are accepted as an inequality that is important to consider in the loan decision. Previously, scholars have made the distinction between "acceptable" vs. "unacceptable" inequalities based on legal precedents between "explainable" and "non-explainable" discrimination [Kamiran and Žliobaitė, 2013] based on Rawlsian philosophy between "relevant" and "irrelevant" features [Rawls, 1999]. For example, income may be considered a "relevant" feature, and gender or race may be considered an "irrelevant" feature. The former should influence the algorithmic decisions, but the latter should not.

Separating what is "relevant" and "irrelevant" is not a simple exercise. In reality, the layers of inequality between two individuals are intertwined, dynamic, and difficult to disentangle from one another. If we consider the layers of inequality in Table 2.4, two individuals may be unequal on several levels – in their level and type of talent, parents' socioeconomic status, behaviour, etc. – that may affect the target outcome of interest, whether it is credit-worthiness, predicted performance at a job, or insurance risk. It is possible that the differences in the observed outcome are attributable to one or more of the above inequalities. Building an algorithm to predict the outcome could result in a faithful – but unwanted and unethical – representation of these inequalities and the resulting replication and perpetuation of the same inequality through decisions informed by its predictions.

Types of inequality	**Examples**	**Variable**
Natural inequality	Disability at birth	Inequality 0
Socioeconomic inequality	Parents'/guardians' assets	Inequality 1
Talent inequality	Intelligence, skills, employment prospects	Inequality 2
Preference inequality	Saving behaviour, cultural prioritisation of values associated with economic opportunities	Inequality 3
Treatment inequality / societal discrimination (external)	Discrimination in job market and education system affecting income stability	Inequality 4

Table 2.4: Layers of inequality affecting the ground truth (partial and indicative) (Adopted from [**Lee** et al., 2021])

The choice of mathematical fairness formalisation determines which inequalities are "unacceptable." Some assume that all disparity in a given outcome metric is unacceptable, while others assume a level playing field [Gajane and Pechenizkiy, 2017], an assumption rarely met in societal challenges. More recent work has taken a more nuanced stance, suggesting that the only features that should contribute to the outcome disparity are those that can be controlled by the individual, emphasising a distinction between the features driven by "effort" vs. "circumstances" [Heidari et al., 2019]. This is derived conceptually from Dworkin's theory of Resource Egalitarianism: no one should end up worse off due to bad luck, but rather, people should be given differentiated economic benefits as a result of their own choices [Dworkin, 1981]. Another paper distinguishes between "benign" disparities and "structural bias" that should be corrected [Binns, 2020].

In §2.2.2.1- 2.2.2.4, we present the limitations of these proposals on how to determine which types of inequalities should be allowed to influence the model's prediction. Then, in §2.3, we mirror the complexity of this decision with the diversity of perspectives in

ethical philosophy that reflect millennia of debate around this topic. Finally, we present a proposal that places fairness considerations in the context of holistic ethical objectives.

2.2.2.1 Why fairness cannot always be assessed through legally protected characteristics

The open source fairness toolkits (discussed further in Chapter 3) designed for ML developers to automatically test algorithms for fairness often refer specifically to **protected or sensitive characteristics** in their assessment of fairness. This is because these characteristics are explicitly defined in non-discrimination laws, and ensuring decision-making is fair along these dimensions is important for legal compliance. For example, the Fairness 360 toolkit defines a protected attribute as one that "partitions a population into groups whose outcomes should have parity. Examples include race, gender, caste, and religion" [Bellamy et al., 2019]. There is limited guidance on under what circumstances two groups should have parity in outcomes, which is important for toolkits intended for usage across domain areas. In addition, how much disparity is acceptable in each use case and for each sub-group of interest? Often, fairness toolkits propose the usage of these demographic features without challenging whether they are relevant to the decision at-hand.

Whether a disparity in fairness metrics between legally protected groups is fair depends on the context. Race and gender may be causally relevant in differential medical diagnosis (e.g. sickle cell anaemia, ovarian cancer) due to the different biological mechanisms in question. If the differences in outcome are causally related to the protected feature, the difference in decisions may be arguably fair. If a man has a higher income than a woman, he may receive a higher credit limit given his higher ability to repay. This is reflected in non-discrimination laws, as there is an exception if the decision-maker can show the policy is a proportionate means of achieving a legitimate aim [Aggarwal, 2018; Wachter et al., 2021; Gillis, 2022].

2.2.2.2 Why fairness cannot always be assessed through relevance vs. irrelevance

Heidari et al. also propose a distinction between relevant vs. irrelevant features [Heidari et al., 2019]. However, Gillis demonstrates through experiments that identifying which features are relevant vs. irrelevant fails to address discrimination concerns because combinations of seemingly relevant inputs may drive disparate outcomes between racial groups [Gillis, 2022]. Lee and Floridi also show that it is possible to use legitimate and relevant features about the loan and predict the applicant's race [**Lee** and Floridi, 2020].

In addition, the notion of "relevance" raises the question of whether a feature that is statistically associated with an outcome should be considered relevant. For example,

in car insurance, even though accident risk varies statistically by gender, pricing and underwriting discrimination based on gender is seen as unfair and faced prohibition in Europe, and there are movements to also ban the usage of racial origin, disability, and sexual orientation [Meyers and Van Hoyweghen, 2018]. However, age is still used in car insurance pricing [Abdou, 2019].

Insurers justify their usage of various features, including age and previously gender, on the basis that they are relevant in assessing people's risk. Based on the technical concept of "actuarial fairness" coined by the neoclassical micro-economist Kenneth Arrow (1921–2017), the price of an insurance policy is considered fair if it is equal for customers with the same risk level [Arrow, 1978]. Otherwise, low-risk groups would be subsidising high-risk groups that have artificially deflated prices.

Importantly, Meyers and Von Hoyweghen argue that there has been a fundamental shift in insurance from actuarial fairness to "behavioural fairness," in which the development in wearable technologies has led to a greater focus on collecting behavioural data for more personalised pricing [Meyers and Van Hoyweghen, 2018]. In other words, there is a shift in conversation not only on what is representative of risk but also whether it can be controlled. However, separating what is a fair feature based on whether it is due to effort or circumstances is also problematic.

2.2.2.3 Why fairness cannot always be assessed through effort vs. circumstances

The suggestion to distinguish between the features driven by "effort" vs. "circumstances" in algorithmic fairness [Heidari et al., 2019] follows the logic of Dworkin's theory of Resource Egalitarianism: no one should end up worse off due to bad luck, but rather, people should be given differentiated economic benefits as a result of their own choices [Dworkin, 1981]. In reality, it is difficult to separate out what is within an individual's *genuine control*. For example, a credit market does not exist in a vacuum; while potential borrowers can improve their creditworthiness to a certain extent, e.g. by building employable skills and establishing a responsible payment history, it is difficult to isolate the features from discrimination in other markets, layers of inequality, and the impact of their personal history.

In addition, some circumstances are valuable features in an algorithm. For example, one may not be in full control of one's income (socioeconomic inequality) or education level (talent inequality), but they are important indicators of credit risk and associated with greater job security. It is challenging to separate the aspects of one's income that are due to life choices within one's control and the aspects that are outside of their control, e.g. due to workplace discrimination or socioeconomic status. As previously discussed, age is still used in car insurance pricing, even though it is not under our control [Meyers

and Van Hoyweghen, 2018].

2.2.2.4 Why fairness cannot always be assessed through the source of inequality

Scholars have also proposed that the *source* of inequality should determine which fairness metric is appropriate for each use case, i.e. whether the outcome disparity is explainable, justifiable, or benign or due to structural discrimination [Kamiran and Žliobaitė, 2013; Binns, 2020]. Binns (2020) suggests group fairness metrics assumes disparities are benign, e.g. the loan approval difference between white and black applicants is solely due to their differences in ability to repay; statistical parity assumes structural bias that requires correction, e.g. historically, black applicants' risk have been inflated due to past discriminatory practices. In reality, there is rarely such a separation [Binns, 2020]. For example, Lee and Floridi (2020) review the literature on U.S. mortgage lending and suggest that there are many structural and statistical factors that lead the lenders to both over-estimate and under-estimate the risk of black borrowers [**Lee** and Floridi, 2020]

Any attempt to isolate the impact of discrimination from the impact of "benign" inequality needs to also consider the intersectional discrimination faced by those already marginalised in society [Crenshaw, 1989], e.g. the inter-connectivity of gender and racial discrimination [Collins, 2002]. The boundary between what is an acceptable representation of existing inequalities and what is due to systematic discrimination and marginalisation of a group is challenging to ascertain.

Fleurbaey (2008) also cautions that "responsibility-sensitive egalitarianism" in welfare economics could be used to hastily justify inequalities and unfairly chastise the "undeserving poor" [Fleurbaey, 2008]. The idea that people should bear the consequences of their choices is not as simple as it seems; it only makes sense when individuals are put in equal conditions of choice. Such an equality is not true in most systems. When one has fewer opportunities than another, one cannot be held fully responsible insofar as one's choice is more constrained. This is further discussed in §2.4 on the lessons from welfare economics.

2.2.2.5 The challenge of defining fairness

The assumed clear and intuitive separation between acceptable and unacceptable inequalities, whether based on their source or the role of luck, rarely exists in real-life models. Therefore, making the distinction on whether a feature is driven by acceptable or unacceptable inequalities is often impractical. In addition, the boundary between what is acceptable vs. unacceptable is more controversial than is often portrayed in the algorithmic fairness literature (especially in computer science). The criteria for desirable equality depend on the philosophical perspective, which is ultimately a subjective judgement. We discuss the

variety of philosophical perspectives in the next section to show the disagreement among ethical philosophers on what is an acceptable inequality.

The decision on the target state—the way it ought to be—is an ethical decision with mathematically inevitable trade-offs between objectives of interest. Heidari et al. dismiss the distinction between relevant vs. irrelevant features in practice as out of scope for their paper: "Determining accountability features and effort-based utility is arguably outside the expertise of computer scientists" [Heidari et al., 2019]. On the contrary, we argue that *model developers must be actively engaged* in the discussion on what layers of inequality should and should not be influencing the model's prediction. The engagement of model developers should be supported by an organisational governance and risk management framework and process. This is aligned to the notion of *anticipatory governance*, which embeds social values, ethics, and public preferences into the scientific research process to shape technologies from an early stage [Guston, 2010]. At minimum, the model developers should be required to transparently disclose the various decisions made that affect the model's fairness to relevant business leader, and they should also help communicate any tensions between the practical and ethical objectives.

The policymakers, regulators, and business leaders all have a role to play in ensuring ML systems we deploy are fair and ethical, including setting government-level and organisation-level policies, laws, and guidelines. However, these policies, laws, and guidelines are implemented in practice by the developers of these systems and necessitates a conversation around the objectives of the ML system in question. This is because such a discussion directly influences not only the model design and feature selection but also the selection of performance metrics. We demonstrate the extent to which the assumptions impact the design decisions in the build process in Chapter 3.

2.3 Lessons from ethical philosophy on (in)equalities

While the consideration of fairness in computer science and machine learning literature is fairly recent [Hutchinson and Mitchell, 2019], ethical philosophers have long debated whether equality is desirable and – if so – what type of equality people should pursue in society. Table 2.5 gives an example of philosophical perspectives and their perceptions of what types of inequality are acceptable. Formal equality of opportunity (EOP), or procedural fairness, posits that all opportunities should be equally open to all applicants (e.g. jobs, loans, etc.) based on a relevant definition of merit. However, in theory, this can be fully satisfied even if it is only a minority segment of a population (e.g. those with family wealth and connections) that have realistic prospects for accessing the opportunity. In other words, as long as the opportunity is theoretically *available*, it is irrelevant whether it is *practically accessible*. It is important to consider the structural inequalities that

Philosophical perspective	**Acceptable inequalities**	**Unacceptable inequalities**
Formal equality of opportunity / procedural fairness [Greenberg, 1987]	Any inequality as long as the opportunity was open to all	Treatment inequality
"Fair equality of opportunity" [Rawls, 1999, 2001]	Natural, talent, and preference inequalities	Socioeconomic, treatment inequalities
Rawlsian EOP + Difference principle [Rawls, 1999]	Natural, talent, and preference inequalities, plus any inequality benefiting the most disadvantaged society members in long-term impact	Socioeconomic, treatment inequalities, except any ineequality benefiting the most disadvantaged society members in long-term impact
Equality of outcome / condition / welfare [Greenberg, 1987]	None - all members should get the exact same outcome	All
Luck egalitarianism [Dworkin, 1981]	Effort-based inequalities (e.g. preference)	Circumstances (e.g. natural inequality)
Equality of freedom / autonomy [Sen, 1992]	Inequality resulting in "genuinely free" choices	Any inequality hindering freedom
Sufficiency / Equality of capability [Walzer, 1983]	Any inequality as long as everyone is above the level of sufficiency	Any resulting in people falling below sufficiency levels
Prioritarianism [Scheffler, 1994; Parfit, 1991]	Any inequality reduction should prioritise resource allocation to those who are worst off	None as long as the worst off are prioritised
Desert [Kagan, 1999, 2014]	Any inequality based on what he/she "deserves"	Any inequality that does not equate to the person's deservingness

Table 2.5: Key philosophical perspectives on inequality (Adapted from [**Lee** et al., 2021])

are deeply embedded into society. For example, racial identity in the U.S. is not simply a personal subjective quality, but rather, an ascribed political category with systemic patterns of social and spatial segregation; ignoring this deep-seated inequality turns a blind eye to the history of disadvantage [Benthall and Haynes, 2019].

The Rawlsian fair EOP goes further to propose that any individuals with the same native talent and ambition should have the same prospects for success, requiring that all competitive advantage (e.g. parental efforts) be offset [Rawls, 1999]. This is at odds with libertarian ideals that assert the value of each person's freedom insofar as there is no harm to another [Mill, 1998], which naturally extends to the right to ownership and capital. Rawls also proposes the Difference Principle as an exception: economic and social inequalities can only be justified if they benefit the most disadvantaged members of society [Rawls, 1999]. These EOP principles are in contrast to the strict equality of outcome, condition, or welfare, which requires an equal distribution regardless of any relevant criteria.

Interestingly, the Maximin rule derived from the Difference Principle is also used in defining "maximin" fairness (a version of EOP) in the distribution of computer network's bandwidth: prioritisation of smaller flows [Denda et al., 2000]. Another fairness definition unique to this domain is *proportional fairness*, which balances between maximising total throughput of the network and at the same time allowing all users at least a minimal level of service [Kelly, 1997].

Luck egalitarians hold that unchosen inequalities must be eliminated [Dworkin, 1981]. Sen and Fleurbaey object on the grounds that luck egalitarians have no principled objection to a society in which, on a background of equal opportunities, some end up in poverty or as the slaves of others [Fleurbaey, 2008]. They argue for a more substantive equality of "autonomy" that includes the full range of individual freedom.

Some have argued that what is important is not relative condition compared to other people, but rather, whether people have enough to have satisfactory life prospects [Walzer, 1983]. Others have shifted the focus on the incremental gain of well-being of those who are worst-off [Parfit, 1991]. Yet others have debated the foundations of desert, or what one deserves corresponding to his or her virtue [Kagan, 2014].

2.3.1 Ethical subjectivity of algorithmic fairness

As such, what types of inequality of outcomes are fair is a philosophical and subjective debate with nuances and complexities insufficiently addressed in existing algorithmic fairness literature. What happens when faithfully representing the world as it is perpetuates an unfair state of affairs? This complicates the objective of ML, which is only 'reliable' insofar as it is trained on data sets that reflects reality. For example, online searches for "CEO" yield mostly images of white men [Van Dam, 2019]. This is reflective of the

existing gender pay gap: in 2019, only 6.6% of Fortune 500 top executives were female, the highest proportion in history [Zillman, 2019], but continuing to under-represent women in search results may perpetuate the bias that CEOs are typically men. Online job postings may show high-income positions to men more frequently than women, reflecting this status quo [Van Dam, 2019], and Amazon reportedly scrapped a recruiting algorithm that preferred men because most of past applicants were men [Dastin, 2018]. Gender bias in online job postings and recruiting algorithms may result in a biased outcome, with men disproportionately having access to jobs. In this instance, some call for the "correction" of the bias to reflect judgements about the way the world *should* be, which is by nature an ethically influenced choice. Indeed, an analysis of EU non-discrimination laws, scholars have argued that in legal cases, a context-specific "legitimate comparator" group that is receiving an unfair advantage should be defined: for example, are married couples equal to same-sex civil partnerships, and are full-time workers equal to part-time workers? [Wachter et al., 2021]. It is not always clear what are "acceptable" inequalities and "unacceptable" inequalities in outcomes. This is mirrored by lab studies of individuals who challenged algorithmic decision-making as unfair because its assumptions do not account for their multiple concepts of fairness [Lee and Baykal, 2017].

As previously stated, and in contrast to past scholars' arguments [Heidari et al., 2019], our position is that computer scientists and model developers cannot completely delegate this consideration to a third party, whether it is the regulator, business leader, or the risk function. Model developers must be engaged in the discussion on what layers of inequality should and should not be influencing the model's prediction in order to inform their decisions on model design, feature selection, and performance metric selection.

Overall, in formalising fairness, the decision-maker should be explicit on (i) which inequalities and biases exist that affect the outcome of interest, and (ii) on which of them should be retained and which of them should be actively corrected. This will be further addressed in §2.5.3, with our proposal for Key Ethics Indicators (KEIs). We next link some of the fairness metrics to the ethical philosophy that inspired them, pointing out the contextual considerations in the ethical philosophy that should be kept in mind alongside the fairness formalisations.

2.3.2 Linking ethical philosophy to algorithmic fairness

Mathematical definitions of fairness, while loosely derived from a notion of egalitarianism, should be calculated while keeping in mind the nuances and context-specificity present in philosophical discourse. Revisiting the fairness metrics from Table 2.2 and Table 2.3, this section will link each metric to the ethical philosophy that inspired it, as well as addressing the gaps between the philosophical work and what is represented in the mathematical formula.

Fairness metric	**Equalising**	**Philosophy**
Maximise total accuracy	N/A	Desert [Kagan, 1999, 2014]
Demographic parity, group fairness, disparate impact [Feldman et al., 2015]	Outcome	Strict egalitarianism (Equality of outcome / condition / welfare) [Greenberg, 1987]
Equal opportunity / false negative error rate balance [Hardt et al., 2016]	FNR	"Fair equality of opportunity" [Rawls, 1999, 2001]
False positive error rate balance / predictive equality [Chouldechova, 2017]	FPR	
Equal odds [Hardt et al., 2016]	TPR, TNR	
Positive predictive parity [Chouldechova, 2017]	PPV	
Positive class balance [Kleinberg et al., 2016]	Average probability of positive class	
Negative class balance [Kleinberg et al., 2016]	Average probability of negative class	
Counterfactual fairness [Kusner et al., 2017]	Prediction in a counterfactual scenario in which the person had a different attribute	David Lewis, cause and effect [Lewis, 1973]
Individual fairness [Dwork et al., 2012]	Outcome for "similar" individuals	Responsibility-sensitive egalitarianism [Fleurbaey, 2008]

Table 2.6: Fairness metrics and their philosophical origins, demonstrating 1) the moral irreconcilability of the metrics and 2) the gap between the original ethical philosophy and how it has been formalised in the metrics (Adopted from [**Lee** et al., 2021])

We now iterate through the definitions in Table 2.6. Accuracy maximisation is prone to biases introduced in the model development lifecycle that may skew the predictions. This is especially problematic if the biases reflect patterns of societal discrimination, leading to "undeserved" outcomes in conflict with the philosophy of desert. Demographic parity is undesirable if there are legitimate rationale behind the unequal outcome (e.g. unequal income). John Rawls has been called "AI's favorite philosopher" [Procaccia, 2021] due to how frequently he is referenced in algorithmic fairness literature (e.g. [Heidari et al., 2019; Dwork et al., 2012; Joseph et al., 2016]). The equal opportunity metric, while it sounds attractively similar to Rawlsian EOP, fails to address discrimination that may already be embedded in the data [Gajane and Pechenizkiy, 2017]. Discrimination may be crystallised in the data set due to biased data collection, biased data labelling, or biased human decisions feeding the system. These biases are introduced throughout the ML development lifecycle and are described in detail in Chapter 3. Rawlsian EOP also assumes that inequalities in native talent and ambition may result in unequal outcomes, which is not addressed in the equalisation of false negative rates. In addition, we previously mentioned the Difference Principle, which posits that the Max-Min social welfare function should also maximise the welfare of those who are worst-off [Rawls, 1999]. Rawls himself explicitly states that "the maximin rule is not, in general, a suitable guide for choices under uncertainty" [Rawls, 1999]. As we discuss in detail in Chapter 5, in building an ML system, there are several layers of uncertainties, many of which are irreducible. Therefore, there are significant gaps between what Rawls envisioned as "equal opportunity" and how it is formalised. A detailed philosophical analysis of the original Rawlsian context and its incompatibility with all but an extremely limited set of AI applications can be found in [Franke, 2021].

Each group fairness metric, including equal odds, positive predictive parity, and positive / negative class balance, requires different assumptions about the gap between the observed space (features) vs. the construct space (unobservable variables): *"if there is structural bias in the decision pipeline, no [group fairness] mechanism can guarantee fairness"* [Friedler et al., 2016]. This is supported in a critique of existing classification parity metrics, in which the authors conclude that "to the extent that error metrics differ across groups, that tells us more about the shapes of the risk distributions than about the quality of decisions" [Corbett-Davies and Goel, 2018]. In many domain areas in which there are concerns over ML fairness, including credit risk and employment, there has often been a documented history of structural and societal discrimination, embedding unfair bias into the training data. Accepting the training data as given embeds an inherent value judgement that replicating any embedded biases is fair [Friedler et al., 2016].

The challenge of individual fairness is how to define "similarity" [Kim et al., 2018]. When the predictive features are also influenced by protected features, measurement of

"similarity" cannot be independent of those protected features. How does a developer separate people's features that affect their creditworthiness from the features of who they are, such as race and gender? As previously discussed, structural inequalities that are deeply embedded into societal institutions should not be ignored [Benthall and Haynes, 2019]. Some scholars have attempted to incorporate active corrections for racial inequality into metrics of similarity [Dwork et al., 2012], but this depends heavily on the assumption that the inequality due to racial discrimination can be isolated from other sources of inequality, which may not be realistic.

While counterfactual fairness metrics provide an elegant abstraction of the algorithm, the causal mechanisms is not well understood in many cases where ML is typically used to model complex relationships in large data sets. Indeed, when the "causal graph" of the mechanisms is unknown, counterfactual fairness is also sensitive to unmeasured confounding variables, which may add additional discriminatory bias [Kilbertus et al., 2020a]. It is also difficult to isolate the impact of one's protected feature, e.g. race, on the outcome, e.g. risk of default, from the remaining features. Confounders are especially difficult to determine for complex models. Critiques of the counterfactual approach have stated, "Even though counterfactuals play an essential part in some causal inferences, their use for questions of algorithmic fairness and social explanations can create more problems than they resolve" [Kasirzadeh and Smart, 2021]. This is because social categories of race and gender should not be subject to quantitative counterfactual manipulation [Kasirzadeh and Smart, 2021]. It is not so simple to posit what the outcome would have been if a person were not of a certain race or gender.

The types of inequalities that are acceptable depends on the context of the model. In all, these metrics do not give information on which layers of inequalities they are attempting to correct, which risks over- or under-correction. Deeper engagement with the ethical assumptions being made in each model is necessary to understand the drivers of the unequal outcomes. Our KEI approach in §2.5 will account for such context-specificity of what inequalities are acceptable.

2.4 Lessons from welfare economics: Consideration of welfare and liberty

In this section, we draw from literature in welfare economics to demonstrate the importance of contextualising *fairness* in the holistic view of *ethics*, which includes welfare and autonomy. By focusing narrowly on the fairness metrics, which quantify the redistribution of the target outcome, a decision-maker may overlook the key considerations of the impact on the stakeholders' welfare and autonomy. Because of the challenge in quantifying the relevant biases and disentangling them from the outcome of interest, correcting for a bias

for the sake of fairness carries the risk of increasing the inaccuracies of the predictions. Referring back to our definition of algorithmic ethics, justice is only one of five dimensions (*beneficence, non-maleficence, autonomy, justice, and explicability*), with fairness as a key principle related to justice. We derive lessons from literature on welfare economics to demonstrate the inter-connectedness of fairness and welfare (*beneficence* and *non-maleficence*) and liberty (*autonomy* and *explicability*). The egalitarian perspective on the relative distribution of resources between individuals and groups must be considered alongside the aggregate impact of an algorithm on the society, as there may be unavoidable trade-offs among them. We later introduce Key Ethics Indicators as a way to explicitly define these trade-offs to enable an informed decision on model design and development.

2.4.1 Welfare in algorithmic ethics: beneficence and non-maleficence

We will continue with the example of credit risk evaluation to argue that fairness should be considered alongside welfare. In attempting to improve a fairness score, a decision-maker may inadvertently forego an algorithm that leaves everyone better-off (beneficence) or may inadvertently harm the sub-group they are attempting to help. Fairness metrics should not be taken at face value without an understanding of how relying on these metrics may affect other ethical objectives. Fairness toolkits that assess fairness in isolation risks misleading the decision-makers by giving the them incomplete information about whether their algorithm meets their ethical objectives.

From a welfare economic standpoint, a notion of fairness necessarily includes a consideration of well-being: from both utilitarian and libertarian perspectives, a fair reward principle maximises the sum total of individual well-being levels while legitimising redistribution that enhances the total outcome of individuals [Fleurbaey, 2008]. This is not necessarily contradictory to the egalitarian perspectives discussed in ethical philosophy. In accordance with the Difference Principle, Rawlsian EOP Max-Min social welfare function should also maximise the welfare of those who are worst-off [Rawls, 1999]. A model that results in financial harm of already-disadvantaged populations fails to meet the Rawlsian EOP criteria, even if the False Negative Rates are equalised as per the mathematical definition. Without consideration of the long-term impact on welfare, the fairness metrics fail to capture the full extent of ethical dilemma embedded in a model selection process.

Accuracy is often considered in trade-off with fairness [Kleinberg et al., 2016], but that accuracy may represent a key ethical principle in beneficence or non-maleficence. For an example of beneficence, a "good" credit risk algorithm would lower the aggregate portfolio risk for the lender, enabling more loans to more people and giving them access to credit that is crucial to upward socioeconomic mobility. As an example of non-maleficence, the false positive rates (i.e. loans that were approved but defaulted) also contains information about whether unaffordable loans are approved. A lender should aim to minimise the

borrower's financial difficulty, given the adverse effects of unaffordable debt on both the market level (causing instability and a "bubble") and for the borrower [Aggarwal, 2018].

There is precedent for prioritising welfare over fairness. In the UK, a joint reinsurance scheme called "Flood Re" was introduced to cap flood insurance premium by Council Tax Band, which some scholars say is inherently unfair [Penning-Rowsell, 2015]. Even though this policy means those with low flood risk are subsidising those with high flood risk, it is prioritising the welfare of those who cannot afford to pay the premium set by their true flood risk. The controversy around Flood Re was around whether this prioritisation is valid and whether the policy reduces the incentive for those living in high flood risk zones to move to a safer area [Penning-Rowsell, 2015]. This also demonstrates the role government may play in controlling the market when actuarial fairness is at odds with the population's welfare.

The ethical principle of non-maleficence may sometimes be in direct conflict with fairness. Adding fairness constraints may end up harming the groups they intended to protect in the long-term [Liu et al., 2018]. In the presence of a feedback loop, we need to consider – not only providing a resource (a loan) to an applicant in a disadvantaged group – but also what happens as a result of that resource being allocated. If the borrower defaults, his/her credit score will decline, potentially precluding the borrower from receiving future loans. It is important to view fairness, not in isolation at a moment in time, but rather, in the context of long-term objectives in promoting the customer's financial well-being. The importance of long-term monitoring will be discussed in Chapter 5.

2.4.2 Liberty in algorithmic ethics: autonomy and explicability

Fairness should also be assessed within the context of how the algorithm affects human *liberty*, a subject in welfare economics that is relevant to the AI ethics principles of *autonomy* and *explicability*. Fleurbaey argues responsibility-sensitive egalitarianism in welfare economics should move away from "responsibility," which may overlook certain people's lack of freedom to choose alternatives, and towards "autonomy" [Fleurbaey, 2008]. For there to be "true" equality, three conditions must be met: 1) a minimum level of autonomy is attained, 2) with a minimum level of variety and quality of options offered, 3) with a minimum decision-making competence [Fleurbaey, 2008]. A comprehensive egalitarian theory of justice is not just about equalising *available* opportunities but also about providing adequate opportunities and making them *accessible*. In reality, defining a minimum level of autonomy can be challenging, but autonomy can be actively considered in the context of an algorithm's potential harms. As per our definition of AI ethics [Floridi and Cowls, 2022], we define *autonomy* as the power to decide, striking a balance between the decision-making power humans retain and that which we delegate to artificial agents. We also define *explicability* as the combination of *intelligibility* (how it works) and *accountability*

(who is responsible for the way it works). It complements the other four principles by helping us understand the good or harm an algorithmic system is actually doing to society, in which ways, and why [Floridi and Cowls, 2022].

2.4.2.1 Autonomy: Liberty

In enforcing some of the stricter fairness conditions, decision-makers should be cognisant of the potential impact this has on human autonomy. For instance, luck egalitarians have no objection in principle to a society in which, on a background of equal opportunities, some end up in poverty or as the slaves of others [Fleurbaey, 2008] – this could violate fundamental human rights to freedom and result in undesirable levels of extreme societal inequality. Intervention is necessary when basic autonomy is at stake, and this should be a constraint on definition of fairness. Fleurbaey argues this is consistent with egalitarian welfare economics, as egalitarians should be concerned not only with equality of opportunities, but also with the content of the opportunities themselves, with freedom as the leading principle in defining responsibility in social justice [Fleurbaey, 2008].

By focusing on equality of opportunities, one may dismiss the differences in preferences as driven by choice and thus irrelevant. However, Fleurbaey argues that the ex post inequalities due to differences in preferences are also a target for intervention on the grounds of improving the range of choices to suit everyone's preferences. If more women prefer lower-paid positions than men, what is problematic is not only the societal and environmental conditioning that questions whether this is a genuine preference, but also the unfair advantage that attaches to these jobs – a differential value of the "menu" of options for women than for men because of their preferences [Fleurbaey, 2008]. Considerations of fairness and the associated policy response must operate at the level of the *menu*, rather than distribution of jobs themselves. This "menu" is a dimension of autonomy that is not captured by the quantitative fairness metrics.

2.4.2.2 Autonomy: Forgiveness

One concept that is not often addressed in algorithmic fairness literature is forgiveness. Fleurbaey argues that the ideal of freedom and autonomy contains the idea of "fresh starts": in absence of cost to others, it is desirable to give people more freedom and a greater array of choices in the future [Fleurbaey, 2008]. This is in conflict with the "unforgiving conception of equality of opportunities" that ties individuals to the consequences of one's choices [Fleurbaey, 2008]. In many countries, lenders are restricted in their access to information about borrowers' past defaults; for example, many delinquencies are removed from U.S. credit reports after seven years [Elul and Gottardi, 2015]. Forcing a lender to ignore information about past behaviour may reduce the accuracy of its default prediction model, and it may be "unfair" by some definitions by putting those who have made more

responsible financial decisions on equal level as those who have not; however, it is widely accepted practice to ensure that one decision does not have a disproportionate impact of limiting one's access to credit for good. A more complete coverage of fairness and justice, therefore, should go beyond redistribution of outcome features and consider the impact on individual welfare, autonomy, and freedom.

2.4.2.3 Autonomy: Vulnerability

Autonomy also cannot be met as an ethical objective when there is a significant asymmetry of power and information between two parties. Contractarian perspectives on fairness assumes two equal entities exchanging one resource for another in rational decision-making [Gauthier, 1986].

Those with limited autonomy include vulnerable people. When an algorithm targets and manipulates those with limited options and recourse, those people do not have the autonomy to enter into the contract, whether or not the contract is fair. Payday loans and check cashing industry in the US targets those who cannot access traditional financial services, often due to their immigration status or long working hours that do not provide a break while a bank is open for business, entrapping the most vulnerable groups into an unbreakable cycle of debt with unaffordable interest rates [Prager et al., 2009]. While the interest rate may not necessarily be unfair (it may be proportional to the likelihood of an individual's repayment), it is ethically undesirable. The same principle applies to marketing insurance products to those with recent bereavement or the sale of complex financial instruments to someone without the capability of understanding their risks.

Another group is those with "thin" files, with a lack of or sparse credit history. This is a fairness issue, as people may be excluded from the data set, either through discrimination or limited accessibility. The resulting skew in the data set is described as "representation bias" in §3.2. However, it is also an issue of autonomy because those who are not well-represented in the data set may be forced to give up more of their privacy to gain the same level of access to products and services [Eubanks, 2018].

There has been a movement to use "alternative data" or non-traditional data sources that do not directly relate to the borrower's ability to repay. One of the most extreme cases is the use of Internet browsing history, location, and payment data to calculate credit risk [Koren, 2016]. The justification is often that this increases financial inclusion for those without alternate means to access credit. However, this requires the lender access to more data from the currently unbanked populations, disproportionately forcing them to give up their privacy, more so than those with existing credit histories. It also exacerbates the risk of discrimination, as the non-traditional data sources are likely intertwined with personal characteristics. Location and social media data are more likely to reveal an individual's race and gender than credit history. While Kenya's poor were among the first to benefit

from digital lending applications, they have led to a predatory cycle of debt the borrowers describe as a new form of slavery, between the endless nudges to borrow, the lenders' control over a vast archive of user data, and ballooning interest payments [Donovan and Park, 2019]. This *double-standard of privacy* between the unbanked and banked violates the equal rights of individuals to privacy and self-determination. While there may be an exchange of access to credit and personal data (e.g. if an individual gives consent to a personality test or access to his/her social media profile), there should be a protection of their right to privacy. This trade-off between not being visible (representative bias) and being seen too much (violation of privacy) has been discussed in past literature on data localisation [Mishra, 2015; Hon et al., 2016] and on data nationalisation [Bellet and Frijters, 2020; Millard, 2013].

Fairness overall must be considered in the context of the impact on individual human rights – going beyond the equality of available opportunities, empowering human freedom and autonomy to ensure *accessibility* of these opportunities. Computer scientists should learn from the welfare economists' consideration of autonomy as a crucial component of egalitarian perspectives on fairness.

2.4.2.4 Explicability

Welfare economics is built on the assumption of rational, free agents, which is shared in Kantian ethical philosophy [Kant and Gregor, 1996]. This has been applied to medical ethics to mandate that a patient be able to make a fully informed decision on whether or not to receive treatment [Eaton, 2004]. Similarly, in algorithmic decision-making, individuals consenting to the usage of their data should fully understand how the data will be used. When humans employ autonomous systems, they cede, at least provisionally, some of their own autonomy (decision-making power) to machines [Floridi and Cowls, 2022]. Respecting human autonomy thus becomes a matter of ensuring that both the decision-making authority and the subject of the decision retain enough autonomy to safeguard their well-being.

In order to incorporate the algorithm into rational decision-making, it is important to understand how the algorithm reached its prediction or recommendation. Due to the relatively limited interpretability of ML, "explainable AI" (xAI) is an ongoing area of research [Xu et al., 2019]. There is often a trade-off between accuracy of an algorithm and its explainability, as complex phenomena are better represented by complex, "black-box" models than simple and interpretable models. This may, in turn, represent a trade-off between explainability (and thus a decision-maker's capability for reasoning) and any beneficence afforded by the increase in accuracy and model performance. In some use cases, e.g. film recommendations, accuracy may outweigh the need for explanations. However, in academic literature on recommender systems, explainability has been emphasised because

it facilitates system design to improve the transparency, persuasiveness, effectiveness, trustworthiness, and system debugging [Zhang et al., 2020]. The explanations may vary based on the target of the explanation, such as the customer, regulator, domain experts, or system developers, depending on the purpose of the explanation [Arya et al., 2019]. One such purpose is to improve the audience's trust of the algorithm. It is important to understand the interplay between an algorithm's explanation and its perceived fairness. There may be a number of possible explanations for any given decision, and the techniques for xAI alone do not detect or correct unfair outcomes. The explanations may help identify potential variables that are driving the unfair outcomes, e.g. if pricing varies for female-dominated professions compared to male-dominated professions, the model may be relying on occupation for its prediction, which acts as a proxy for gender.

Overall, we have shown that while fairness formalisations may provide a simple methodology for model developers to incorporate metrics relevant to equalisation of outcomes between groups and individuals, they do not provide a holistic view of the important debates on what fairness means. In ethical philosophy, the debate hinges on what types of inequalities each scholar believes is acceptable or unacceptable. Fairness in ML should similarly be considered as a fundamentally subjective topic. We then pointed out that the narrow definition of fairness may not consider the long-term and big-picture ethical goals. Drawing from welfare economics, we emphasise the importance of considering welfare and autonomy alongside fairness to understand any competing objectives or trade-offs that may exist in designing an ML system.

2.5 Proposed method: Key Ethics Indicators

In the final section of this chapter, we propose an approach (adapted from our paper [**Lee** et al., 2021]) that moves away from attempts to define fairness mathematically, and instead, gain a more holistic view of the ethical considerations of a model. This is a *process* of making *explicit* all the ethical considerations in a use case, rather than an empirical method. This approach is aligned to the more recent works in computer science that call on researchers and practitioners to *explicitly* document data collection processes, worldviews, and value assumptions [Friedler et al., 2016]. We introduce this to be a part of the business process for an organisation developing ML systems. Due to the subjectivity of fairness metrics, it may be challenging for the decision-makers to select one over another for a system. Rather than these general metrics, decision-makers should create a customised measurement of what "fair" looks like in each model. In practice, the decision-makers would be include both technical developers and non-technical stakeholders, and there may be an approval process, e.g. for the board to review the developer's definition of success. Wider communities, such as affected customers or marginalised groups, may be

surveyed throughout this process to gain their views on what they view as fair. This is especially important in global organisations, in which different cultures, jurisidictions, and communities may have different views on prioritisation of their values. In addition, fairness should not be considered in isolation from the related ethical goals. The interaction between fairness and other values - e.g. welfare, autonomy, and explicability - should be accounted for in this analysis.

Despite claims to the contrary [Heidari et al., 2019], the roles and responsibilities of an engineer are necessarily intertwined with the role of the expert or business stakeholder, as the ethical and practical valuations of what "success" looks like in the model directly influences the algorithm design, build, and testing (to be discussed in Chapter 3). It is important to have active engagement from the beginning between the developer and the subject matter expert to try to understand which inequalities should or should not influence the outcome. This process requires engagement from all relevant parties, including the business owner and the technical owner, with potential input from regulators, customers, communities affected by the ML systems, and legal experts.

Relying solely on the out-of-the-box fairness definitions as implemented in fairness toolkits would fail to capture nuanced ethical trade-offs. There are opportunities for open source communities, technology companies, and other practitioners to contribute to the toolkits to improve them; we will discuss this in Chapter 4.

For a decision-maker, it is important to devise customised success metrics specific to the context of each model, which, as we described, involves considering welfare (beneficence, non-maleficence), autonomy, fairness, and explicability. This can be done through the following process (also visualised in Figure 2.1):

1. Define "success" from an ethical perspective. What is the benefit of a more accurate algorithm to the consumer, to society, and to the system? What are the potential harms of false positives and false negatives? Are there any fundamental rights at stake?

2. Identify the layers of inequality that are affecting the differences in outcome.

3. Identify the layers of bias.

4. Devise an appropriate mitigation strategy. Note this may require changes to data collection mechanism or to existing processes, rather than a technical solution.

5. Operationalise these objectives into quantifiable metrics, build multiple models and calculate the trade-offs between the objectives covering all ethical and practical dimensions.

6. Select the model that best reflects the decision-maker's values and relative prioritisation of objectives. [**Lee** et al., 2021]

We now elaborate each of these steps, in turn.

Figure 2.1: Proposed KEI process

2.5.1 Define success

For each use case, there are unique considerations on what is considered a "successful" model, which are unlikely to be captured in a single mathematical formula. In credit risk evaluation, for example, three key objectives from ethical, regulatory, and practical standpoints are: 1) allocative efficiency: a more accurate assessment of loan affordability protects both the lender and the customer from expensive and harmful default; 2) distributional fairness: increasing access to credit to disadvantaged borrowers, including "thin-file" borrowers and minority groups; 3) autonomy: both increased scope of harm due to identity theft and security risk and due to the effects of ubiquitous data collection on privacy [Aggarwal, 2018]. There are multiple motivations to consider 2 and 3, especially legal and regulatory compliance with non-discrimination and privacy laws and reputational risk of unfairness. Perception of (un)fairness can impact companies' profitability; for instance, an experiment found that on average, people move twice as much money away from banks that use algorithms in loan application decisions when told that they draw on proxy data for race and gender or social media data [Chonaire and Meer, 2020].

A successful credit risk model would achieve all three objectives, though in reality, there may be trade-offs among them. In algorithmic hiring, success metrics may include employee performance, increased overall diversity among employees and in leadership, and employee satisfaction with the role. It is important to identify all the objectives of interest, such that any trade-offs between them may be easily identified, allowing for a more holistic view of algorithmic ethics.

2.5.1.1 Special focus: impact on fundamental human rights and vulnerable populations

The system's potential impact must be assessed at this stage, especially in relation to fundamental human rights and especially algorithmic decisions affecting vulnerable populations. The use of AI systems can have negative impacts on fundamental rights (including those relating to discrimination, privacy, or expression). It is therefore important that those using AI consider impact on such rights. The Data Protection Impact Assessment (DPIA), as enshrined in the **General Data Protection Regulation (GDPR)** [European Union], already tasks organisations with considering the fundamental rights risks of their

undertakings as they relate to rights and freedoms. However, assessing fundamental rights impacts can be challenging for private organisations, as fundamental rights – originally intended to protect the rights and freedoms of citizens against a powerful state – are broad and abstract in nature. Nevertheless, an organisation's assessment (be it public or private) requires a context-specific impact assessment to these rights. Table 2.7 presents a non-exhaustive list of rights within the clusters that organisations must consider in an impact assessment. See our paper [Jenssen et al., 2022] proposing a practical framework to assist organisations in undertaking fundamental rights impact assessments for details on how FRIA can be performed as a part of DPIA in real-life settings. We give examples of risk factors and potential scale of impact on fundamental human rights.

Cluster of rights	**Rights included**
Privacy rights	personal autonomy, private life, sanctity of the home, physical and psychological privacy, communication secrecy, development of one's identity, relational privacy, data protection, right to (not)hold a conviction or a belief
Expressional rights	freedom of expression, artistic expression, commercial expression, freedom to receive and impart information, press freedom, commercial expression, right to assemble, right to vote
Procedural rights	right to motivation, right of access to a court, right to a fair trial, equal access to documentation in court proceedings, right to adversarial proceedings, right to an effective legal remedy
Equality rights	right to equal application of the law to everyone to whom the law applies, prohibition of arbitrariness, prohibition of direct (intentional) or indirect (unintended) discrimination
Socio-economic rights	equal access to health care, equal access to affordable housing, equal access to education, equal access to social benefits

Table 2.7: Overview of fundamental rights relevant in AI context; all of these require assessment as part of a DPIA

2.5.2 Identify sources of inequality

As previously discussed, due to the complex and entangled sources of inequalities and bias affecting an algorithm, there is no simple mathematical solution to unfairness. It is important to understand what types of inequality are acceptable vs. unacceptable in each use case. Table 2.4 presented different layers of inequality. Forcing the decision-maker to look beyond the legally protected characteristics to identify the inequalities that are acceptable and relevant and those that are not helps better identify the sub-groups that are at risk of discrimination.

We previously claimed that computer scientists and model developers should actively engage in the discussion on what layers of inequality should and should not be influencing the model's prediction in order to inform their decisions in the development process.

An accountability mechanism, such as the assignment of roles and responsibilities, is important; we have addressed how to embed risk management in the AI development lifecycle in our paper [Lee et al., 2020], and we have mapped tools and techniques for unfair bias mitigation to a standard organisational risk management lifecycle [**Lee** and Singh, 2021b], which will be discussed in Chapter 3. We have also proposed a framework for "reviewability" to ensure the logs, reporting, and audit trail are fit for purpose for understanding the AI models [Cobbe et al., 2021], to be further discussed in Chapter 5. In these papers, we emphasise the need for ethical principles to be operationalised into practice and embedded into organisational processes, ensuring that the right stakeholders are involved at the appropriate stage and that the accountability and responsibility of each ethical risk is clear.

2.5.3 Identify sources of bias

In addition to the inequalities discussed above, there may be biases in the model development lifecycle that exacerbate the existing inequalities between two groups. The challenge is that in many cases, the patterns associated with the target outcome are also associated with one's identity, including race and gender. In Chapter 3, we will propose a practical tool in identifying unintended biases in these six categories; here, we give a brief overview in the context of KEIs.

Suresh and Guttag (2020) have recently grouped these types of biases into 6 categories: historical, representation, measurement, aggregation, evaluation, and deployment. Historical bias refers to past discrimination and inequalities, and the remaining five biases, displayed in Table 2.8, align to the phases of the model development lifecycle (data collection, feature selection, model build, model evaluation, and productionisation) that may inaccurately skew the predictions. By understanding the type of bias that exists, the developer can identify the phase in which the bias was introduced, allowing him or her to design a targeted mitigation strategy for each bias type.

Table 2.8 gives examples of racial discrimination in lending processes to demonstrate each type of bias. Crucially, they point out that effective bias mitigation addresses the bias at each stage of the lifecycle, including non-technical interventions. For example, bias introduced through the data collection process may require a change in marketing strategy. This stage of KEI approach enables bias to be addressed in each stage of the development pipeline in which it was introduced.

2.5.4 Design mitigation strategies

The mitigation strategy depends on whether we believe the inequalities in Table 2.4 and the biases in Table 2.8 need to be actively corrected to adjust for inequalities and bias.

Types of bias	Examples	Variable
Representation bias	Limited marketing and outreach in high-minority neighborhoods	Bias 0
Measurement bias	Unequal treatment in the lending process associated with race leads to mis-measurement of risk factors	Bias 1
Aggregation bias	There may be a difference in default frequency distribution between racial groups, which is poorly represented by a single model	Bias 2
Evaluation bias	The accuracy and precision metrics in default prediction vary across racial groups (e.g. lower confidence in predictions for minority borrowers)	Bias 3
Deployment bias	True outcome only known for accepted loans and unknown for denied loans	Bias 4

Table 2.8: Layers of bias resulting in inaccurate predictions (partial and indicative)

It is important to understand the type of bias and in which stage of the lifecycle it was introduced in order to address it. In Chapter 4, we will discuss testing and mitigation; however, we will provide a brief overview below in the context of KEIs.

There have been existing methods proposed for **pre-processing**, removing bias from the data before the algorithm build, **in-processing**, building an algorithm with bias-related constraints, and **post-processing**, adjusting the output predictions of an algorithm (See: §4.2). However, these methods presume that inequalities in Table 2.4 and the biases in Table 2.8 are known and can be quantified and surgically removed. How do we isolate the impact of talent and preference inequalities on income from the impact of discrimination? The attempt to "repair" the proxies to remove the racial bias has been shown to be impractical and ineffective when the predictors are correlated to the protected characteristic; even strong co-variates are often legitimate factors for decisions [Corbett-Davies and Goel, 2018].

Often, the solution to these biases is not technical because their sources are not inherent in the technique. Instead of looking for a mathematical solution, there may be productive ways of counteracting these biases with changes to the process and strategy. Examples are shown in Table 2.9.

While the mitigation strategies are important, they are unlikely to provide a complete solution to the problem of algorithmic bias and fairness. That is because—unlike the assumptions underlying fairness formalisations—it is often not feasible to mathematically measure and surgically remove unfair bias from a model, which is affected by inequalities and biases that are deeply entrenched in society and in the data. Selbst et al. 2019 have argued that "technical interventions [are] ineffective, inaccurate, and sometimes dangerously misguided when they enter the societal context that surrounds decision-

Types of bias	Variable	Example action
Treatment inequality / societal discrimination (external)	Inequality 4	Identify a new feature to estimate income volatility associated with race
Representation bias	Bias 0	Change in marketing and outreach strategy to include more high-minority neighbourhoods
Measurement bias	Bias 1	Employee training on subconscious bias, standardised practice on which loan types are recommended based on pre-specified relevant criteria
Deployment bias	Bias 2	Continuous monitoring and analysis of whether the decision boundary between rejection and acceptance is appropriate

Table 2.9: Possible actions to counteract biases (*partial and indicative)

making systems" and highlight the need to focus on the process as well as the technical in a system.

Legal scholars have argued that traditional approach of scrutinising the inputs to a model is no longer effective due to the rising model complexity. Using Fair Lending law as an example, Gillis demonstrates that identifying which features are relevant vs. irrelevant fails to address discrimination concerns because combinations of seemingly relevant inputs may drive disparate outcomes between racial group [Gillis, 2022]. Rather than focusing on identifying and justifying inputs and policies that drive disparities, Gillis argues, it is important to shift to an *outcome-focused* analysis of whether a model leads to impermissible outcomes [Gillis, 2022]. Similarly, Lee and Floridi have proposed an approach to assess whether the outcome of a model is desirable [**Lee** and Floridi, 2020]. For a more comprehensive analysis of whether a model meets the stakeholders' ethical criteria, it is important to look beyond the inputs and the designer's intent and assess the long-term and holistic outcome. Overall, designing the mitigation strategies would then enable the calculation of any *residual* risk after they have been implemented. These residual risks should be considered alongside the potential benefits in the next stage, in which we operationalise the risks and benefits into KEIs.

2.5.5 Operationalise KEIs and calculate trade-offs between KEIs

Once "success" for a model has been defined at a high-level, the next step is to operationalise the ethical principles such that they are measurable. Similarly to how a company may define a set of quantifiable values to gauge its achievements using Key Performance Indicators (KPIs), there should be outcome-based, quantifiable statements from an ethical standpoint: Key Ethics Indicators (KEI), enabling developers to manage and track to what extent each model is meeting the stated objectives. This may include positive benefits of

the model as well as potential harm (residual risks).

For example, Lee and Floridi estimate the impact of each default risk prediction algorithm on financial inclusion and on loan access for black borrowers [**Lee** and Floridi, 2020]. They operationalise financial inclusion as the total expected value of loans under each model and minority loan access as the loan denial rate of black applicants under each model. In Figure 2.2 replicated from our past work, they calculate the trade-offs between the two objectives for five algorithms, providing actionable insights for all stakeholders on the relative success of each model.

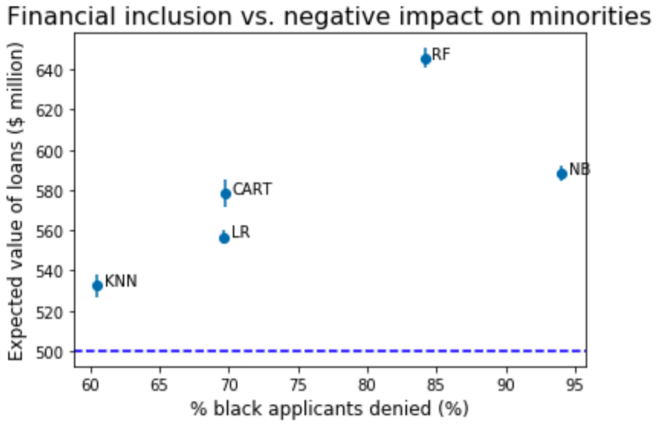


Figure 2.2: Replicated from Lee and Floridi (2020): Trade-off analysis

Context-specific KEIs can be developed for each use case. For example, in algorithmic hiring, employee satisfaction with a role may be estimated by attrition rates and employee tenure, employee performance may be measured through their annual review process, and diversity may be calculated across gender, university, region, age group, and race, depending on each organisation's objectives and values. It is possible that not all considerations are identified in the beginning, and stakeholders may disagree with one another on how to select and prioritise KEIs. However, the important step is making explicit the ethical objectives in each use case. This would help decision-makers justify the use of any algorithm, rendering the trade-offs more transparent and reviewable. The decision-makers may add, remove, or change the KEIs as their understanding of the objectives evolve through the model development process. In the long-term, explicit discussion of KEIs in each use case can lead to the establishment of industry standards, informing best practices, policy design, and regulatory activity.

Once the KEIs, risks, and mitigation strategies have been identified, the stakeholders can assess whether the potential benefits of the model outweigh the residual risks after the mitigation has been taken into account. This should inform the initial go/no-go decision at the design phase.

2.5.6 Select a model and provide justifications

While the model selection and justification will be discussed in Chapter 3 and Chapter 4, this will be briefly show how the trade-off analysis between these pre-established KEIs makes the ethical considerations clear and actionable to the decision-maker. For example, in Figure 2.2, Lee and Floridi conclude that Random forest is better in absolute terms (in both financial inclusion and impact on minorities) than Naïve Bayes, but the decision is more ambiguous between CART and LR: while CART is more accurate and results in greater financial inclusion (equivalent of $15.6 million of loans, or 103 median-value loans), CART results in a 3.8 percentage points increase in denial rates for black loan applicants compared to LR. This quantifies the concrete stakes to the decision-maker who may decide on the model that is most suited to his or her priorities, customised to each use case.

One of the key benefits of the outcome-driven KEI trade-off analysis is that it provides interpretable and actionable insights into the decision-maker's values, which is especially important for complex machine learning algorithms in which the exact mechanism may not be transparent or interpretable. This could also provide valuable justification to the regulator on why a certain model was seen as preferable to all other reasonable alternatives. This may also help reduce the hesitation among decision-makers around the use of machine learning models due to their non-transparent risks, if the analysis shows they are superior to traditional rules-based models in meeting each of the KEIs. Suitable records of the decisions must be kept, ensuring the model and its design are *reviewable* [Cobbe et al., 2021]. This is further discussed in 5 on how to review and monitor KEIs in online learning settings.

2.6 Key chapter takeaways

Mathematical fairness definitions have been implemented into technical toolkits without locating their implications in overall algorithmic ethics. One of our contributions is to derive lessons from *ethical philosophy* and from *welfare economics* on what are the *contextual considerations* that are important in assessing an algorithm's ethics beyond what can be captured in a mathematical formula. For example, we refer to the debate in ethical philosophy on what constitutes acceptable vs. unacceptable inequalities. We also relate to the explicit consideration in welfare economics of welfare and liberty, which are associated with algorithmic ethics principles of beneficence, non-maleficence, autonomy, and explicability. Over-reliance on fairness metrics would capture only one dimension of an algorithm's ethical impact.

As a step forward, our second contribution is the proposal of a generalised "Key Ethics Indicator" (KEI) approach that *explicitly* forces a consideration of the ethical objectives, aligning to the contextual features that we have drawn out as important

in ethical philosophy and welfare economics literature. This is in contrast to previous ethical frameworks that have been permutations of similar principles of beneficence, non-maleficence, justice, autonomy, and explicability. The KEI approach requires translation of the technical fairness metrics into real-world implications, such as the potential impact on the customers or users of the ML system. It also enables reviewable record-keeping of any trade-offs and alternative models that were considered and why one model was chosen over another. By actively accounting for the inequalities and biases, the KEI approach more closely aligns to the lessons from ethical philosophy than the technical fairness metrics. By considering fairness alongside welfare and autonomy, the KEI approach aligns to lessons from welfare economics. In certain cases, especially in high-impact public sector models, the KEIs may be used as a transparency mechanism on how decisions are made and released to the public for any contestation or debate. This is aligned to the ethos of responsible innovation that is founded on a substantive "inclusive reflection and deliberative democracy."

Often, the discomfort with the use of ML to make decisions derives from the tension between the opportunity provided by algorithms that can more accurately predict an outcome and the risk of systematically reinforcing existing biases in the data and the risk of undermining human autonomy [Lee et al., 2020]. On the other hand, unlike human subconscious biases, the ethical impact of machine predictions can be systematically audited, debated, and improved [Kleinberg et al., 2018]. By understanding the holistic ethical considerations of each algorithmic decision-making process using KEIs, decision-makers can be better informed about the value judgements, assumptions, and consequences of their algorithmic design, opening up the conversations with regulators and with society on what is an ethical decision.

In addition to identifying these risks in KEI, it is also important to test and interrogate the KEI considerations. This will be discussed in Chapter 4. Once the algorithm has passed the go/no-go decision through the assessment of the risks and benefits, the developer moves to the build phase. Once the model has been deemed to have greater benefits than risks in the design phase, in a business process, often the developer would gain permission to proceed with the build. However, by the iterative nature of an ML development process, the model design often evolves throughout its build, with potential trade-offs changing among the KEIs. The KEI approach proposes the building of multiple models, calculating the trade-offs, and selecting the model that best reflects the decision-maker's values. These KEIs are then monitored. We will discuss the build phase in the next chapter.

Chapter 3

ML Build: identifying and measuring unintended unfair biases

Introduction

In building an ML model, the developer makes many decisions that may affect the model's fairness. In the KEI approach in Chapter 2, Step 3 was to identify the layers of bias in the system. While some may be apparent from the design phase, such as the representation bias that exists in the data set that has already been collected, the developer building the ML model has to be mindful of other potential **unintended biases** that are being introduced through his/her modelling decisions that may result in a disproportionately negative impact on minority, under-represented, vulnerable, or excluded communities. These decisions include data collection mechanism (if the developer decides to collect data instead of or in addition to using a pre-existing data set), feature selection / engineering, decision on how to measure the outcome, model selection, and selection of performance metrics.

In the Introduction, we discussed the incentives for organisations to combat fairness-related risks. In this chapter, we again focus on *unintended* biases under the assumption that it is within the ML developer's interest to reduce undesired discriminatory outcomes. It is plausible that there are malicious developers intending to unfairly discriminate and mask their intention throughout their ML development process by making the ML model look fair. However, these adversarial actors are out of scope for this thesis because they have an entirely different incentive structure. We aim to provide guidance to the well-intentioned ML developers on how to identify unintended unfair biases.

While frameworks for identifying the risks of harm due to unintended biases have been proposed [Suresh and Guttag, 2021], these biases had not been operationalised into practical tools to assist industry practitioners until our paper [**Lee** and Singh, 2021b]. In this chapter, we introduce a bias identification methodology and questionnaire first

proposed in [**Lee** and Singh, 2021b], illustrating its application through a real-world practitioner-led use case. We validate the need and usefulness of the questionnaire through a survey of industry practitioners, which provides insights into their practical requirements and preferences. Our results indicate that such a questionnaire is helpful for proactively uncovering unexpected bias concerns, particularly where it is easy to integrate into existing processes, and facilitates communication with non-technical stakeholders. A questionnaire that helps the developer contextualise the bias risk on the potential impact is crucial to its applicability.

Ultimately, the effective end-to-end management of ML risks requires a more targeted identification of potential harm and its sources throughout the model build lifecycle, so that appropriate mitigation strategies can be formulated. Towards this, our questionnaire provides a practical means to assist practitioners in identifying bias-related risks.

3.1 How a system's definition as AI and ADM provides only partial information about its risk

Before we discuss the bias risk identification in ML build process, it is important to understand these risks are not inherent in the *technique* of AI and ML. That is, a system without ML components may also have fairness-related risks. There is a growing range of guidance that relates to governance of technical systems that specifically reference and target AI and ML in areas including privacy and data protection, fundamental rights, and ethical considerations. These AI and ML guidance documents are issued by regulators [Information Commissioner's Office, 2017], governments [Government of the Netherlands; OECD.AI Policy Observatory], legislative bodies [European Parliament], and international organisations [European Commission Independent High Level Expert Group on Artificial Intelligence; European Commission, a; Council of Europe Commissioner for Human Rights]. While these documents understandably advise on the risks of ML-driven systems that are newly and increasingly adopted across industries, the framing around AI and ML could give organisations the mistaken impression that AI systems are exceptional and higher-risk, requiring separate attention to non-AI systems. This is problematic for two reasons. First, the term 'AI' is inherently ambiguous and open to interpretation, as discussed in §1.1. Second, it is difficult to tease out the nuances in the overlaps and 'grey areas' between ML techniques and/or automated decision-making (ADM) processes in any system. Therefore, it is important to be clear on these terminologies and distinguish between guidance unique to ML and that which is applicable across all systems.

Adapted from our Chapter in the 2022 *European Data Protection Handbook* [**Lee** et al., 2022], in this section, we argue that – given the nuances masked by such terminologies – organisations should adopt risk-oriented approaches to identify system risks that extend

beyond technology classification as AI or non-AI. The publication is intended for an audience beyond the technical developers; therefore, this section is framed to be accessible to legal scholars and policymakers. Guidance and recommendations that target typical concerns regarding technology specifics, such as the technique (AI) or the degree of automation of an organisation's processes (ADM), will often only partially capture a system's risk profile [Cobbe et al., 2021]. The proliferation of guidance specific to AI may, for instance, give rise to the mistaken assumption that all AI systems are higher risk than non-AI and require exceptional and separate risk management. A more holistic assessment of system risk is needed, rather than based on some 'top-down' categorisation of the technologies employed. Thus, while we frame our thesis around AI and ML, we refer wherever appropriate to contextual risks, emphasising that the mitigation should be proportionate to these risks. This is a part of our critique of existing approaches that are not tailored to the context-specific risk levels.

A system's categorisation as an 'ADM system using AI' or 'ADM system without AI' gives only partial information about each system's risk profile. Given the overlaps among the three terminologies, these are best visualised in a Venn diagram among three sets: AI, profiling, and ADM. Considering Figure 3.1, AI is associated with *what a system uses*, profiling is associated with *what a system does*, and ADM is associated with *what a system is used to do*.

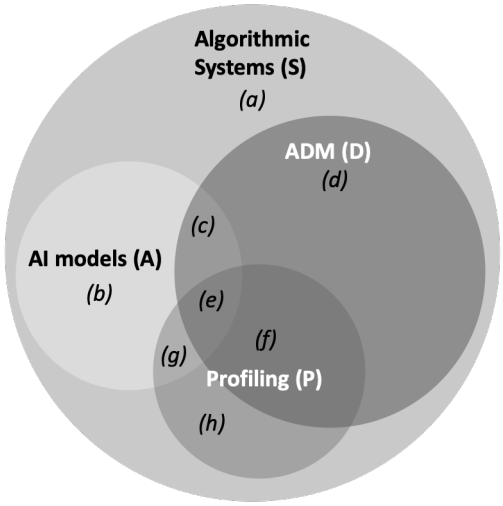


Figure 3.1: Legal definition of AI, ADM, and profiling

A system may involve any combination of AI *techniques*, ADM *processes*, and profiling *activities*. Whether AI is being used for ADM, i.e. a system using AI to make automated decisions, depends on 1) its purpose and 2) the extent to which human intervention is applied. A retail chatbot, for example, may interact with a customer without human input but provide only informational support, e.g. whether a particular product is in stock (label *(b)*). The chatbot's intent prediction algorithm arguably would not be considered ADM.

An ML algorithm that analyses the text in job applications to automatically reject under-qualified candidates would use AI in an ADM process for the purpose of profiling people (label *(e)*) due to its use of ML techniques and its processing of personal data to make automated decisions with respect to individuals. However, an ML algorithm trained on the same data to tag text in the CV relevant to an advertised role could be considered as using AI for profiling but not necessarily used for ADM (label *(g)*) [European Commission, b]. This is because of the role that the algorithm is playing in the decision-making process; in the former example, the algorithm is the decisive agent, while in the latter, the algorithm merely facilitates and accelerates the human decision-making on who is hired.

Conversely, not all ADM would involve AI. Consider a scorecard to apply pre-defined criteria to screen out unqualified candidates; for example, having obtained a postgraduate degree adds 5 points to the total score, and each year of job experience adds 2 points to the score. This is an ADM system that involves profiling but is rules-based in its nature and therefore does not use AI (label *(f)*). As a real-life example, the UK's EU citizens settlement scheme [Tomlinson] involved solely ADM with legally significant effects that did not use either profiling or AI (label *(d)*). The algorithm automatically accepted an application if the information on matched records held by other organisations and if the applicant met a 5-year residency criteria. Because it is not evaluating personal characteristics but rather, classifying people by known information, it is not profiling [European Commission, b].

An algorithmic system may neither use AI nor be used for ADM, such as a rules-based program that highlights keywords in an application prior to human review (label *(a)*). Examples of systems that use AI for ADM but not profiling (label *(c)*) include those in algorithmic trading that do not involve personal information.

Profiling—automated personal data processing to evaluate personal aspects—may also be a part of ADM process or involve AI as a technique. GDPR characterises profiling as sometimes being part of solely automated decision-making (see, e.g., Article 22(1), Recital 71) on which other decisions may in turn be based (e.g. Article 35(3)(a), Recital 73), but profiling is itself not necessarily automated decision-making (e.g. Recital 24 or 70) [European Commission, c]. Even when it *is a part of* solely automated decision-making, it isn't necessarily solely automated decision-making *with legal or similarly significant effects* (Article 22). For example, general profiling may be performed without links to any decisions, e.g. high-level insights into the customer base distribution (label *(h)*). This could include ML techniques, e.g. clustering customers using unsupervised machine learning techniques into categories (label *(g)*). If decisions are made off the basis of these insights, such as being used for individualised pricing, the system would then be a part of ADM (label *(e)*).

The differences in terminological categories *(a-h)* represented in this figure are complicated and open to interpretation. Due to the inherent ambiguity and nuance around these

terminologies, the *framing of guidance around these terms may inadvertently mislead or confuse as to the document's intended or applicable scope*. While the guidance documents target their recommendations around 'AI' and/or 'ADM', the organisations must ensure they appropriately interpret to what extent the guidance is applicable to non-AI and/or non-ADM systems with similar risk profiles. Organisational governance should be framed around specific risks associated with a system, rather than depending solely on a system's classification. This would facilitate a more targeted documentation and logging for greater auditability, testing, and reviewability.

Whether a system uses AI is only partially relevant in assessing its risk; therefore, applying the guidance solely to what an organisation considers AI would be misaligned to the true risk profile of each system. For example, regulators responsible for GDPR enforcement have released guidance framed on AI. In the UK, the ICO released a report [Information Commissioner's Office, 2017] specifically on the implications of AI, big data, and machine learning on the enforcement of GDPR. While the guidance describes itself as generally applicable, its framing around AI, ML, and big data makes it ambiguous to what extent the recommendations apply to systems that do not employ AI. The guidance justifies its focus on AI by claiming that it presents distinct challenges: the use of ML algorithms, the opacity of the processing, the tendency to collect 'all the data,' the re-purposing of the data, and the use of new types of data. However, these considerations are not unique to AI; similar risk factors that may be present in non-AI algorithms. A hiring scorecard may have hundreds of criteria with a complex logic flow, using third-party data sets and applicants' social media profiles. The fact that it is rules-based and not using machine learning techniques does not detract from the potential regulatory risks, including GDPR and EU non-discrimination laws.

There are challenges in any algorithmic system, but not all of them can be attributed to the technique (AI vs. non-AI) selected. Accordingly, the ICO rightly emphasises in the aforementioned report that it is not relevant how an organisation defines AI: "If you are processing this data in the context of statistical models and using those models to make predictions about people, this guidance will be relevant to you regardless of whether you classify those activities as ML (or AI)" [Information Commissioner's Office, 2017]. Some characteristics of AI may require different considerations of risk mitigation, such as for uncertainty in the feedback loop (discussed in Chapter 5), but the the overall risk level is not solely determined by the technique.

All algorithmic systems, AI or non-AI, may be under scrutiny for potential regulatory violations, and organisations are expected to ensure that the governance processes are fit for purpose for each algorithm, in accordance with GDPR [European Union] Articles 24, 25, 32, 35 and other applicable law. However, the term "AI" is overloaded, and the challenge lies in interpreting to what extent the guidance is applicable to non-AI

systems, especially in cases where it has some of the characteristics (e.g. use of alternative, non-traditional data sets) that are typical of an AI. A more holistic risk assessment is needed, and organisations should take into consideration a broader set of risk factors than the selected technology. Further examples and case studies demonstrating this can be found in our *European Data Protection Handbook* chapter [**Lee** et al., 2022].

It is important for organisational governance processes to entail a more holistic assessment of system risk, rather than relying solely on 'top-down' categorisations of the technologies employed. A 'bottom-up' risk identification process enables a more effective identification of appropriate controls and mitigation strategies. Therefore, this thesis concerns itself with less with generalisations on the risks of AI associated with fairness but rather more with approaches that can be tailored to each context, tackling the source of each risk. This is the foundation for our proposed questionnaire in the next section.

3.2 Unintended bias risks in ML build lifecycle

As shown in §3.1, it is advisable to identify potential risks in the end-to-end system, regardless of whether it uses AI. In this section, we address a typical supervised ML development lifecycle, but this can be applicable to non-ML algorithms, such as rules-based scorecards. Instead of focusing on ML techniques alone, we assess the socio-technical system of a developmental lifecycle, including the humans and processes embedded in the pipeline. Thus, we identify the types of bias risks that exist in each stage of a model build process. We use a case study of ML in insurance fraud prediction to validate our proposed questionnaire.

Scholars have proposed mathematical methods (e.g. [Dwork et al., 2012; Hardt et al., 2016; Kusner et al., 2017]) to formalise and test for a particular definition of "fairness" in ML. As discussed in Chapter 2, these definitions can be incompatible with one another [Kleinberg et al., 2016; Pleiss et al., 2017], prompting work distinguishing between them [Verma and Rubin, 2018; Narayanan, 2018]. These techniques assume that fairness can be mathematically operationalised, a view often criticised as overlooking the societal and historical contexts [Green and Hu, 2018; Selbst et al., 2019]. While these mathematical fairness tests may identify *whether and how* a model is "unfair," they do not answer *why*. This makes it difficult to identify mitigation strategies or translate the bias into real-world potential impact. Different metrics provide different answers related to a system's "fairness." In Chapter 1 we observed that these definitions give little information or guarantee on model fairness and argued that there should be a more systematic method for identifying the risk of unfairness – a difficult task where there are competing definitions. In particular, in past user studies, practitioners have claimed they struggle with "explicitly considering biases and 'blind spots' that may be present in the humans embedded throughout the ML

development pipeline, such as crowd-workers or user study participants" [Holstein et al., 2019]. These would not be identified in the mathematical tests and require a qualitative identification.

Referring back to our definition of **bias** in Chapter 1, instead of attempting to define a contextually complex concept such as *fairness*, recent work has also suggested it may be more helpful to identify potential *biases* that skew the outcome in unintended, undesirable ways. Suresh and Guttag (2021) in particular have noted that while downstream harms are often blamed on "biased data," they arise from distinct categories of biases that each aligns to an ML development process [Suresh and Guttag, 2021]. In each stage of model development, there are decisions made that could result in skewing of the outcome in a way that is discriminatory against certain sub-groups, e.g. in data collection and labelling methods, feature engineering, etc. Specifically, Suresh and Guttag (2021) define six categories of biases that may cause unintended harm [Suresh and Guttag, 2021]:

1. **Historical bias**: misalignment between the world as-is and the values or objectives required from the ML model;

2. **Representation bias**: under-representation or failure for a population to generalise for groups in population;

3. **Measurement bias**: choosing and utilising features/labels that are noisy proxies for real-world quantities;

4. **Aggregation bias**: inappropriate combination of heterogeneous, distinct groups into a single model;

5. **Evaluation bias**: use of inappropriate performance metrics or the testing / external benchmark that does not represent the entire population; and

6. **Deployment bias**: inappropriate use or interpretation of model in a live environment.

This echoes similar work on categorising undesired biases [Mehrabi et al., 2021; Olteanu et al., 2019]. The ML development lifecycle involves a series of decisions from evaluation methodology to model selection that can lead to unwanted effects (illustrated in Figure 3.2). As such, instead of "fairness," we refer to **unintended biases** in this chapter with an eye to any aspects of the data, model, and processes in the model build decisions that may result in negative impact, especially on previously marginalised groups. In other words, while the ethical considerations have been identified in the design phase, in the build phase, the developer should be mindful of any decisions that may skew the outcomes in unintended ways during the development.

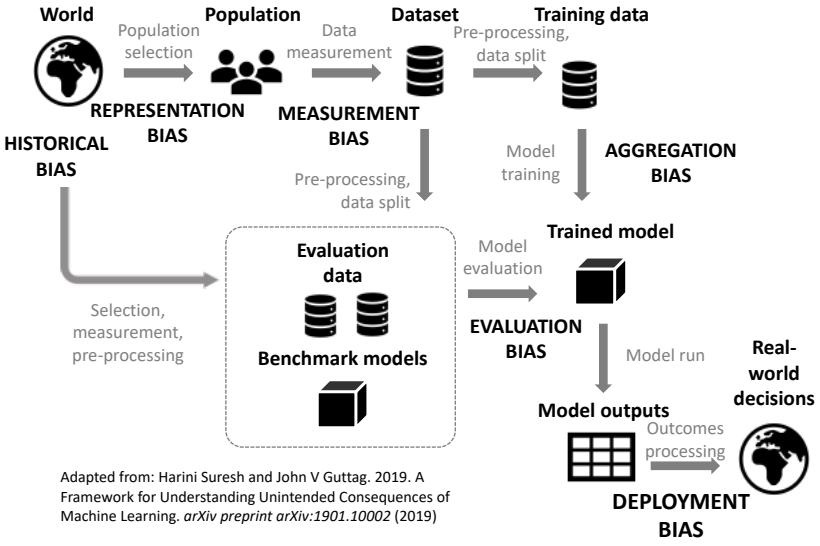


Figure 3.2: Bias in ML development lifecycle

Scholars have found industry practitioners still struggle with challenges of unintended biases. Past studies of practitioner needs have found a significant gap between the methods introduced in research for managing biases and the institutional realities [Veale et al., 2018]. Practitioner approaches to managing the risks of potentially unfair biases is often reactive—focused on addressing customer complaints—rather than proactive, and practitioners are uncertain on how to identify the potential bias risks in their particular context and domain area [Holstein et al., 2019]. Such difficulties for practitioners remain despite the emergence of fairness toolkits (to be discussed in Chapter 4), in part due to the tools' limited coverage of ML lifecycle and the confusion on how such methods integrate with organisational processes [**Lee** and Singh, 2021a]. While Suresh and Guttag (2021) discuss two case studies of unintended biases, they do not provide any generalised method to identify them [Suresh and Guttag, 2021]. Only once bias risk is identified can it be evaluated, quantified, and mitigated, and it represents a significant gap in implemented methods.

3.3 Proposed tool: Risk identification questionnaire

Practitioners believe existing tools and approaches are insufficient in providing clear, targeted processes for identifying the risks of unintended biases and the appropriate mitigation strategies [Veale et al., 2018; Holstein et al., 2019; **Lee** and Singh, 2021a]. While frameworks have been proposed for risk identification of unintended biases, in the context of this thesis, each bias is defined in an operational sense: how it may manifest itself in practical settings [**Lee** and Singh, 2021b] by introducing a risk identification questionnaire that helps to detect the potential risks for each type of bias in each phase of the ML development lifecycle. In other words, the bias framework was broken down into

Questionnaire section	Bias type
A. Background information	N/A - context
B. Design	Historical / external bias
C. Data collection	Representation bias
D. Feature engineering	Measurement bias
E. Model build and training	Aggregation bias
F. Model evaluation	Evaluation bias
G. Model productionisation & monitoring	Deployment bias

Table 3.1: Questionnaire structure

its component parts and translated into practical ways in which the biases may manifest in the lifecycle. In the context of this thesis, it supplements the KEI approach step 3, which is to "identify sources of bias." While some biases may be apparent in the design phase, others may be introduced or discovered during the build.

This marks a departure from the checklist approach by Madaio (2020), which is not aligned to any bias type frameworks and describes *activities* rather than questions helping to elucidate bias risks [Madaio et al., 2020]. For example, the checklist items include "solicit inputs and concerns on system vision" and "undertake user testing" with some example considerations [Madaio et al., 2020]. By contrast, our questionnaire is not intended as an activity checklist, but rather, aims to *engage the developer by walking them through each way in which unintended bias may manifest itself in ML development*, which would help identify a more targeted mitigation activity than a generalised activity list. A checklist is a set of steps to tick off, which are general activities, while a questionnaire prompts to developer to a potential issue related to bias.

The questionnaire aims to provide a starting point for extension and customisation to a particular domain or scenario. After reading the framework by Suresh and Guttag (2021), I have operationalised the concepts into question formats. Future work could further adapt the questionnaire and develop additional guidance on how it may be applied in different contexts. We use a case study on fraud detection in §3.4 to illustrate its usefulness as a general tool. The risk identification process may be carried out internally (through different organisational teams) by the model development team with input from others, such as by legal risk teams, by the internal audit/model validation team, and/or externally for an independent third-party assessment of the ethical risks of the model. The subsequent risk analysis stages, which should be addressed in future work, may be used to assess the trade-offs in the model and justify its usage to key stakeholders, both internal (e.g. board) and external (e.g. customers, regulators).

As Table 3.1 shows, the structure of the questionnaire (outlined below) aligns to the bias framework of Suresh and Guttag (2021) [Suresh and Guttag, 2021] in Fig. 3.2.

Section (A) establishes the context. This mirrors the KEI approach closely, highlighting the various legal, practical, and ethical considerations and the potential positive and

negative impact of the model. The subsequent sections ask probing questions for each stage of the model development lifecycle.

Section (B) considers historical and external biases in the design phase, asking the users to consider the acceptability of existing inequalities in the world they aim to model, such as the presence of any history of discrimination in their domain area.

Section (C) considers the potential for representation bias in the data collection process, challenging the users on any subjectivity of recorded features and the possibility the sample in the data set is not representative of the target population.

Section (D) helps identify potential measurement bias introduced in the feature engineering and selection process, such as any differences in data quality, measurement methods, or the presence of proxies of sensitive features.

Section (E) challenges whether aggregation bias may be introduced in the model build and training, where heterogeneous groups and mechanisms are being improperly accounted for in one model.

Section (F) looks for evaluation bias in the model evaluation process, including the trade-off identification we discussed in the KEI approach and the any disparity in model performance across groups. Finally,

Section (G) looks at the deployment biases that may be introduced in the post-production monitoring process of a model, such as any skewed feedback mechanisms and external changes that may introduce biases into the model re-training.

Answering "yes" to the prompting questions indicates a risk of bias in that phase, prompting its analysis, impact assessment, and mitigation, to be covered in subsequent Chapters 4, 5, and 6. The full questionnaire can be found in the Appendix A. A small sample of the questionnaire is displayed in Fig. 3.3.

3.4 Real-world case study: Questionnaire applied to insurance fraud

We will now walk through a real-world case study to demonstrate the types of bias risks that are identified. The purpose of the case study is to demonstrate how unforeseen biases can be surfaced through this questionnaire. Since this case study, there have been important developments to support practitioners in risk management of AI. A notable example is the National Institute of Standards and Technology (NIST) AI Risk Management Framework, designed as a part of the U.S strategy for AI [Tabassi, 2023].

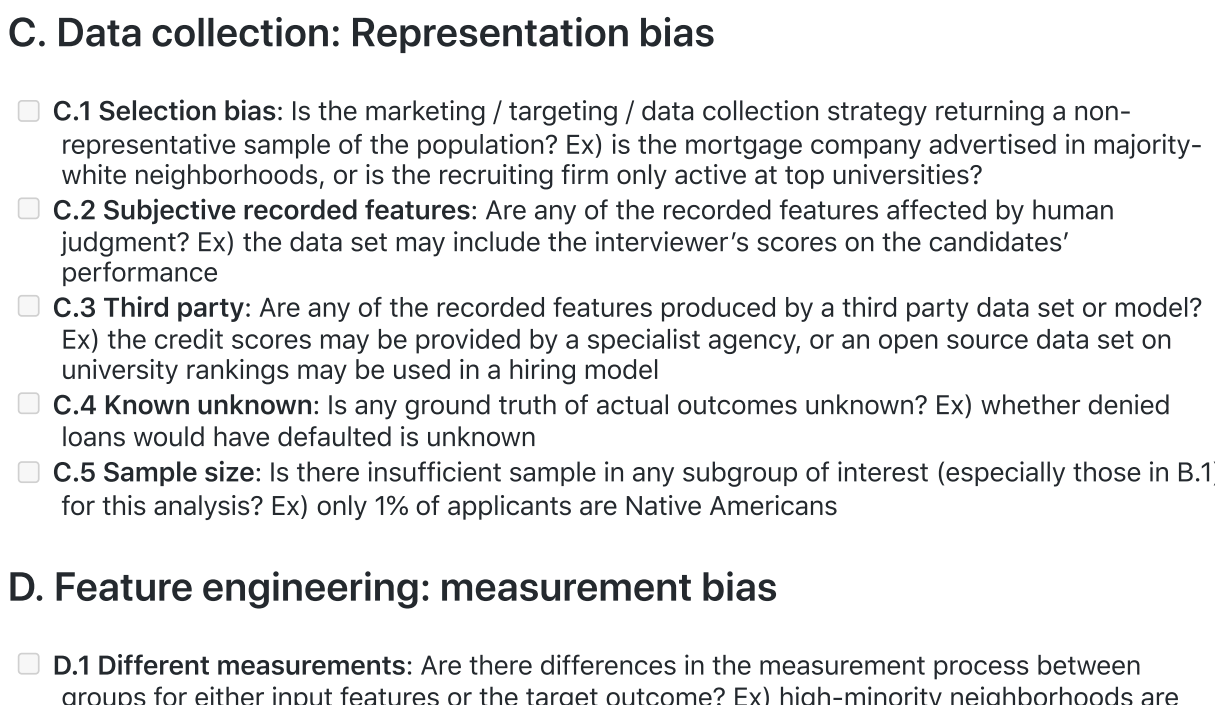


Figure 3.3: Sample snapshot of the questionnaire

The case study was based on a 1:1 walk-through of the questionnaire with the developer of a fraud prediction model for an insurance company (hereinafter referred to as "developer"). All potentially identifying information on the individual, model, and company is withheld to preserve confidentiality. The answers are summarised and paraphrased for conciseness, but all content is contributed by the model developer without our assistance or consultation. Therefore, what is discussed in this section reflects the views of the developer and are not our own.

3.4.1 (A) Background information

The questionnaire begins by probing on the potential positive and negative impacts of the model. It calls for operationalisation of ethical objectives, which we covered in §2.5.1. The first step in defining "Key Ethics Indicators" (KEIs) (§2.5.1) is defining "success" that span ethical, regulatory, and practical perspectives. This is important to contextualise the potential impact of unintended biases and prioritise the types of biases that are the highest risk. In the example of insurance fraud detection, the developer was able to identify the potential benefits and harms in the system. Higher true positive rates in identifying fraud would reduce claim costs, enabling cheaper insurance premiums and reducing money available to criminals. Higher true negative rates would ensure genuinely honest claims are paid more quickly with fewer intrusive processes. Conversely, high false positive rates can make honest claimants feel persecuted, who may withdraw their claims, while potentially

appearing to the regulators and customers as a deliberate bar to making claims. There is also potential representational harm: fraud classification may be taken as an indication of criminality and re-enforce historical and societal discrimination. High false positive rates among marginalised groups may exacerbate this perception and disproportionately affect their financial well-being. Once the metrics (false positive and false negative rates) are translated into real-life implications, they are imbued with practical significance to inform decisions made throughout the build process.

3.4.2 (B) Design: historical/external bias

This section of the questionnaire addresses historical bias, which is relevant to ML models when the world as faithfully represented in the training data does not align with the ideal "target" world. If there is documented historical discrimination in the domain area, e.g. history of racial discrimination in employment, then training a model on the data would replicate this bias. This is associated with the step in Key Ethics Indicators in which we identify the sources of both "acceptable" and "unacceptable" inequalities in the real world that may be represented in the given data set (§2.5.2).

The developer suggested that the identification of potential criminal acts is regularly accused of racial or faith-based biases. Regarding which types of inequalities are a justifiable source of differences in model outcome, the developer answered the only demographic information that may be considered is the preferences of an individual, i.e. choice to deceive by action or inaction, or pattern of behaviour that show they are likely to commit fraud. There is no evidence socioeconomic background is a potential indicator of fraud risk on its own, but the developer stated that it may be justifiable in combination, e.g. a low-income claimant for an expensive watch. Race, gender, disability, age, national origin, talent/education level, personality traits, culture, and discrimination in related markets (e.g. employment) should not play a role on their own in affecting the prediction of fraud risk.

3.4.3 (C) Data collection: representation bias

Collection methodologies can skew how the data set represents the ground truth. The developer identified four representation biases by using the questionnaire.

First, the majority of data used in the insurance claim fraud risk assessment is entered into the system by a claim handler, which may result in subconscious judgement being embedded into the input data. For example, the developer suggested a possibility that claimants who do not speak English well could, e.g. due to miscommunication with the claim handler, result in a different quality of data.

Second, some features in the data are collected by suppliers or specialists as a part of

the claims process. Third party data sets may have their own sets of selection biases that may not be representative of the company's client base.

Third, any claim that has not been investigated is labelled as honest, and there is a general assumption that a significant percentage of fraud is missed because it is not flagged in human or machine screening. These "unknown unknowns" suggest that some actual outcomes are mislabelled, and any models built on previously investigated claims would find similar cases of fraud and be unable to detect the non-obvious cases that are incorrectly recorded as honest.

Fourth, it was noted that the proven fraud rate in insurance claims "rarely exceeds 2% and significantly lower in some business lines." It is especially challenging for a model to identify patterns when there is an insufficient sample of any subgroups of interest represented in the full data set.

3.4.4 (D) Feature engineering: measurement bias

Measurement bias may be introduced in the feature engineering process if there are differences in the measurement process between groups for either input features or the target outcome. The developer identified several measurement biases through the questionnaire. Fraud models can rely on features engineered by the model developer based on fraud intelligence or histories, which could themselves be biased and affected by developer judgement. There is also a risk of **proxies** in measurement: any attempts to locate geographical patterns of fraud could create unintended correlations with certain national or racial groups. (Note the issue of proxies is discussed in further detail in §3.6.) The target outcome measure is also imperfect: a model can only identify claims for further investigation, which is not the same as confirmed fraud. As mentioned, it is assumed that there are cases of fraud that are missed by both the model and the investigator.

3.4.5 (E) Model build and training: aggregation bias

In searching for potential biases in model build processes, the questionnaire attempts to uncover aggregation biases, which occur when populations are heterogeneous in a way such that a single model cannot account for all subgroups. The developer noted that, because there is no single type of fraud, a good detection model must identify which of the many possible fraud scenarios may have occurred and flag it appropriately to the investigation team. The model may be improperly aggregating different types of fraud with different causal mechanisms.

3.4.6 (F) Model evaluation: evaluation bias

The questionnaire then considers whether the model is over-fitting to a particular metric, such as accuracy. The developer emphasised that the relative importance of false positive and negative results can vary according to the business appetites and claim types. A false positive can lead to a sub-optimal customer experience due to delays in the company's payment of a legitimate claim. A false negative involves a financial loss to the company due to an unidentified fraudulent claim. While both metrics are considered, the developer noted the core metric for a fraud model is whether a claim is appropriate for further investigation (true positive rates), which can emphasise the flagging of outliers rather than genuinely fraudulent claims.

3.4.7 (G) Model productionisation and monitoring: deployment bias

The questionnaire also probes on potential biases in the model once deployed, as an ML model is often a part of a complex socio-technical system with inter-connected models or embedded in human processes. The developer answered that fraud models feed human investigators, who flag any claims which were not correctly marked for investigation. Investigators' biases may continue to reinforce any biases in the model as the key feedback mechanism. If there are any external changes that may affect the model, the team manually reviews and implements any model changes.

Overall, the developer stated that while some of these biases were known limitations, such as fraud being a rare event and the difference between potential and confirmed fraud, answering the questionnaire helped systematically list them and consider their cumulative impact. The developer noted that it makes the mitigation strategies clear for each type of potential bias. These views were echoed in the following survey.

3.5 Survey results on questionnaire's effectiveness

We conducted an online survey of industry practitioners to (i) better understand practical requirements for risk identification materials in real-life use cases, and (ii) validate the effectiveness and usability of the questionnaire on a larger variety of scenarios and domains. The study passed our departmental ethical review process and used Qualtrics survey software. It was anonymous and did not ask for any identifying information, including name, company, or contact details. We emailed the survey link to direct contacts, as well as advertising it on online communities related to data science and analytics, e.g. those on meet-up, Facebook, reddit, and LinkedIn groups. We also encouraged sharing of the survey link to anyone working in data science and analytics. Of the 105 people

Roles	N	%
Academic	10	13%
Business lead	8	10%
Data scientist	29	37%
ML engineer	8	10%
Software developer	1	1%
Technical lead	11	14%
Other	11	14%

Table 3.2: Survey demographics: roles

who started the survey, 78 (74%) of the respondents completed at least one section and 29 (28%) completed the entire survey. Summary statistics on their roles can be found in Table 3.2. Due to our contacts being primarily in the UK, most (69%, n=53) were from the UK, with the remainder in the US (18%, n=14), Belgium (3%, n=2), India (3%, n=2), and 1 respondent each from Canada, France, Hong Kong, Ireland, Netherlands, and Singapore. The limitations of our sample and our research design are discussed in §3.7.

All practitioners were given a link to the full questionnaire to read through it with their own use cases in mind and answer whether the questionnaire was helpful. We structured the survey into the following four sections: (1) Demographics, (2) Importance of different characteristics of bias assessment, (3) To what extent the questionnaire meets these criteria, and (4) the questionnaire's usability. In (2), we asked for ratings on various criteria of a risk assessment questionnaire from "Extremely important" to "Not at all important", with probing questions to explain their answers. In (3), we asked how the questionnaire meets the criteria from "Strongly agree" to "Strongly disagree." In (4) we used the standard System Usability Scale (SUS) [Brooke, 1996b] to measure usability.

We also asked the practitioners, if they are comfortable doing so, to share the bias-related challenge they have faced in their work, in order to contextualise their answers to the survey. 16 respondents chose to share the details of their model, which included a diverse set, e.g. recruiting, sales forecasting, genetic disease prediction, facial recognition, appointment no-shows, and content moderation.

The survey aimed to not only validate the questionnaire but contextualise how the practitioners may use these types of tools in their work. We now report on our findings, highlighting takeaways on practitioners' needs and preferences.

3.5.1 Uncovering unexpected biases

Our results show that bias is clearly of concern to the respondents. Our survey confirmed that 90% of practitioners believe the "ability to proactively diagnose unexpected issue(s)" related to biases is extremely/very important. 86% of them agree that our proposed questionnaire meets this need. Practitioners commented in the free-text answers that

Statement	% who believe it is extremely or very important	% who agree that the questionnaire meets this need
Ability to proactively diagnose unexpected bias issue(s)	90%	86%
Ease of integration into existing processes	83%	62%
Facilitating communication with non-technical stakeholders	81%	79%
Identifying potential mitigative actions	78%	59%

Table 3.3: Survey findings: practitioner needs that are met by the bias risk questionnaire. See §3.7.1 for some of the explanations

Statement	% yes
My organisation would use this questionnaire	65%
I think I would like to use this questionnaire frequently	59%

Table 3.4: Survey findings on future usage by practitioners

the "breakdown of different types of biases," "clear structure," "standardising model assessment," and "concrete concepts" are the most helpful aspects of the questionnaire, helping practitioners "think about bias in a systematic way." One practitioner responded it was "bringing up points that wouldn't have occurred to me," and another said it "allowed me to consider a broader range of impact points that may affect my model's bias than I would have otherwise been aware of." More broadly than the risk diagnosis, the questionnaire was found to enable greater familiarity with the model. 77% believe "better understanding of model risk" is extremely/very important important, with 83% agreeing the questionnaire helps them achieve this goal.

3.5.2 Ease of integration into existing processes

The practitioners reported the importance of a bias tool's "ease of integration into existing processes" (83% extremely/very important). Regarding the questionnaire, 62% agreed that our proposal fulfilled this aim, with 24% neutral and 14% disagreeing. In answering whether the practitioner's organisation would use the questionnaire, 65% said "yes", while 35% said "no." A few who disagreed explained it was not directly relevant to their work, one stating it would require domain-specific modifications and adaptations. Others who agreed answered in free-text that it "can be integrated straight away" and "would fit in well with our existing risk management, documentation, and approval processes." This shows how the questionnaire can be broadly applied across domain areas as it is in its general state, as well as representing a starting point for others to adapt to their use cases where further customisation is required.

3.5.3 Facilitating communication with non-technical stakeholders

One feature that the practitioners ranked of high importance is "facilitating communication with non-technical stakeholders" (81% extremely/very important). 79% agreed the questionnaire is helpful in this regard. One practitioner commented that the questionnaire provides "a good set of examples, which can help educate on the need for such a process." Another noted it is "an accessible step-by-step document that can outline bias points that could be understood by my target audience."

3.5.4 Usability and the challenge of designing a simple tool for a challenging problem

We aimed to measure the usability of the questionnaire to understand its accessibility and user-friendliness, in addition to its function in bias identification. To this end, we used System Usability Scale (SUS) [Bangor et al., 2009]. SUS provides a standardised measurement to compare the toolkits to supplement the topic-specific questions, as the toolkits aim at both developers and higher-level practitioners (see above) and can inform non-technical stakeholders. While SUS is most often used for interface design, it has been used in other contexts as well [Bangor et al., 2009], and the questions were asked here to provide a standard basis of measurement for its usability.

The average SUS score of our questionnaire out of 100 was 65.3, with standard deviation (sd) of 17.9. A study of 1,000 SUS surveys showed that "poor" average SUS score is 35.7 (sd 12.6), "OK" is 50.9 (sd 13.8), and "good" is 71.4 (sd 11.6) [Bangor et al., 2009]. While SUS scores may vary by tool type, this provides an intuitive reference point for our questionnaire, which would fall between "good" and "OK" based on the score alone. However, this should be viewed in the context of the tool being a questionnaire without a user interface or user interaction, for which the practitioners had brief, one-off exposure, in which its primary concern was its content rather than its design. The most important point is the content's usefulness for practitioners. In the SUS survey, 59% of the interviewees agreed with: "I think I would like to use this questionnaire frequently." Since the writing of this questionnaire, it has been adopted by several large, global organisations. Importantly, however, it was clear that some respondents wanted the questionnaire to do more – address the analysis, mitigation, and impact assessment, which were beyond the scope of a questionnaire's purpose. However, this shows the importance of this thesis' purpose: to present an end-to-end solution that covers the entire lifecycle.

One point of disagreement regarding the questionnaire's usability was its scope as a qualitative process, despite a quantitative approach being incompatible with bias risk identification. While some welcomed the qualitative design (e.g. "ethical qualitative

assessment... should be the precursor to any machine learning project"), three of the respondents objected to its lack of quantifiable metrics in the free-text comments. Three respondents suggested there should be a "scoring system," with one observing, "I just feel engineers like a quantitative approach." Another practitioner claimed to be in favour of the questionnaire but was unsure whether it could be adopted in their organisation because "model development seems to be quite quick atm [at the moment] with a focus on quantitative processes. I think it would be hard to get engineers to agree on a qualitative outcome." The complex social nuances and implications of model bias depend heavily on each context and would be difficult to quantify [Selbst et al., 2019; Green and Hu, 2018]. Weighting each risk in a scoring system would also only be feasible once biases and their impact are understood in the further analysis stages.

While around half (50%) agreed the questionnaire was "short and focused on high-risk points," others challenged the length, impatient with the more in-depth and contextual bias consideration. One would prefer "a 10 bullet point questionnaire." Another said "I prefer 2-steps (post-processing) in order to make it simpler," referring to "de-biasing" mitigation techniques (e.g. Kamiran (2012)) that correct model outputs to equalise a given metric. This does not align with the questionnaire's intent, which aims to identify the *sources* of biases, including those human-/process-oriented, that may not be addressed through technical means. It demonstrates what Selbst et al. (2019) call a "solutionism trap" in "fair-ML" communities: the failure to recognise that the best solution may not always involve technology. While these other approaches (e.g. de-biasing) may fit in as part of a broader mitigation strategy, they should not be treated as a panacea for all bias risks. This shows the difficulty in designing a simple tool for a complex problem. Due to the contextual nature of fairness and bias, it is difficult to prescribe generic mitigation methods. Nevertheless, the questionnaire was well-received among practitioners from various backgrounds and disciplines as being useful in fulfilling their needs.

3.5.5 Helping identify potential mitigation

Practitioners expressed concerns around mitigation, with 78% answering that "identifying potential mitigative actions" were extremely/very important 59% agreed the questionnaire meets this need. Again, note that determining mitigation strategies is not in scope for the questionnaire, yet practitioners found the questionnaire to be helpful in pointing them in the right direction for mitigation. One commented, "the point of each question and what needs to be done to mitigate the bias are clear." Another noted "I particularly like the way the questionnaire links specific questions that are easy to reason about and answer to underlying real-world issues. This gives the user both an understanding of problems that can arise and a sense of the concrete ways they manifest." Of those that disagreed, one said it should be then tied to providing advice on "how to identify bias at a technical level,"

which is not a part of the identification process and should be addressed in subsequent phases of the development lifecycle.

3.6 Illustrating the challenges of identifying a mitigation strategy: Measurement bias due to proxies

One type of bias that has become a focus in policy discourse on algorithmic fairness is measurement bias due to **proxies**. "Proxy variables" are mentioned in the U.S. White House report on big data [of the President and Podesta, 2014]. The U.S. Department of Housing & Urban Development proposed a change to its interpretation of the disparate impact doctrine: "any material part on factors that are substitutes or close proxies for protected classes under the Fair Housing Act" [Willis et al., 2019]. The controversies in current events have often focused on potential proxy discrimination. For example, a U.K. investigation found insurers reportedly quoted higher premiums if a driver's name is "Mohammed" compared to the quote for "John" with all other data entry being identical [Leo, 2018]. With big data analytics, algorithms are incorporating non-traditional types of data, such as Internet browsing history to predict credit risk [Koren, 2016]. In this section, we will provide an overview of proxy-specific measurement biases with example case studies and the potential mitigative actions.

3.6.1 Legality of proxies

The exact definitions of a "proxy" and "proxy discrimination" have been contested [Tschantz, 2022; Slaughter et al., 2020]. From a legal standpoint, proxy discrimination involves a seemingly neutral practice that disproportionately harms members of a protected class [Prince and Schwarcz, 2019]. Proxies are a feature or a combination of features that encode information about **protected or sensitive characteristics**. This may be intentional or unintentional. A prime historical example of this practice is "redlining," or using zip code as a proxy for race. This not only limited the financial future of black mortgage applicants but also led to significant health inequalities, effects of which still persists today with higher prevalence of cancer, asthma, and poor mental health in previously "red-lined" neighbourhoods [Nardone et al., 2020]. However, for algorithms that rely on correlations between variables, discrimination may not be intentional but go undetected. The nuances of whether such unintended algorithmic discrimination is legally considered *disparate impact* and/or *indirect discrimination* depends on the jurisdiction and the context [Adams-Prassl et al., 2022]. While such legal analysis is out of scope for this thesis, regardless of its classification, it is still mis-aligned to non-discrimination law.

Even without malicious intent, with the prevalence of AI and ML, Prince and Schwarcz

(2019) argued that "AI can and will use training data to derive less intuitive proxies for directly predictive characteristics when they are deprived of direct data," posing a "substantial threat to the normative underpinnings of these anti-discrimination regimes" [Prince and Schwarcz, 2019]. This is because ML relies on correlations in large data sets, many of which may be associated with the outcome of interest, such as loan risk, but also may be associated with who the customers are, including race and gender. In credit risk, for example, given the abundance of non-traditional, "alternative" data being used as a proxy for creditworthiness, there is rising concern over the risk of discrimination against potential borrowers based on legally protected characteristics, such as race and gender. In our paper, we demonstrated using peer-to-peer lending data that the presence of spelling errors is associated with a higher likelihood of loan default [**Lee** and Singh, 2021c]. In this case, spelling errors are intended as proxies of borrower's personality traits previously found to be associated with repayment success (e.g. conscientiousness or carelessness). However, spelling errors may also act as a proxy for dyslexia and for national origin if the borrower's first language is not English.

There are legal exceptions on when a proxy may be used. Given Supreme Court decisions in the UK [Lowenthal, 2017] and the US [Baum et al., 2015], even if a variable is correlated to a protected feature, there may be reasonable grounds to use it if the differences are crucial to a legitimate business requirement. This decision boundary may shift depending on the context. The drivers of decision-making in providing essential products, such as current account, car insurance, or mortgage, may be subject to higher scrutiny than the rationale for offering premium credit cards. This ensures that the decision to include features correlated to protected characteristic is carefully considered within the context of the regulated domain and the potential impact on the customers. For example, Fuster et al. (2022) have shown that there is a difference in income distributions between racial groups [Fuster et al., 2022]. As income has a reasonably inferential relationship to credit risk, it cannot be considered as a simple proxy attribute of protected characteristics.

3.6.2 Proxies: to include or remove?

The obvious mitigation for a proxy is to exclude it, such as removing zip code from a loan application. Care must be taken to ensure the removal of the proxy does not disproportionately reduce predictive performance of the model, as this may have practical and ethical implications described in §2.5. The proxy may be replaced with more precise indicators of outcome. For example, for the peer-to-peer lending data, we propose further isolating the types of spelling errors that are more likely to be associated with carelessness compared to the types that are associated with dyslexia or a lack of familiarity with the English language [**Lee** and Singh, 2021c]. In practice, isolating more precise outcome indicators from the proxies may not be possible [Andrus et al., 2021] but should be

considered as a potential mitigation strategy.

Some scholars have proposed methods for "de-biasing" the data set (**pre-processing**) to remove bias from the proxies [Feldman et al., 2015]. These automated mitigation techniques are discussed and critiqued in §4.2. Intuitively, these methods are impractical when there is not a neat separation between the "legitimate" variations between groups that is necessary to consider in a model, e.g. difference in income for credit risk, from the variations due to undesirable inequalities, e.g. discrimination in the job market. When the proxies are tightly tied to the model development process due to their high association with the outcome, they are challenging to remove or "de-bias" without affecting the integrity of the model. Sometimes, the outcome variable themselves that the model is trying to predict may be a proxy, such as arrest rates in the US which are associated with race [Dressel and Farid, 2018]. Removing these variations may result in reductions in accuracy, which may harm the groups it is trying to help [Liu et al., 2018].

Given the contextuality of whether proxies are problematic, we have previously proposed an approach to determine whether or not an input variable should be used in a model. Figure 3.4, replicated from [Lee et al., 2020], visualises a possible decision boundary for whether or not an input variable should be used in a model, based on its role as a potential proxy for a protected characteristic. This decision boundary may shift depending on the context. The drivers of decision-making in providing essential products and services, such as current account, car insurance, or mortgage, may be subject to higher scrutiny than the rationale for offering premium or niche products and services. This ensures that the decision to include features correlated to protected characteristic is carefully considered within the context of the regulated domain and the potential impact on the customers.

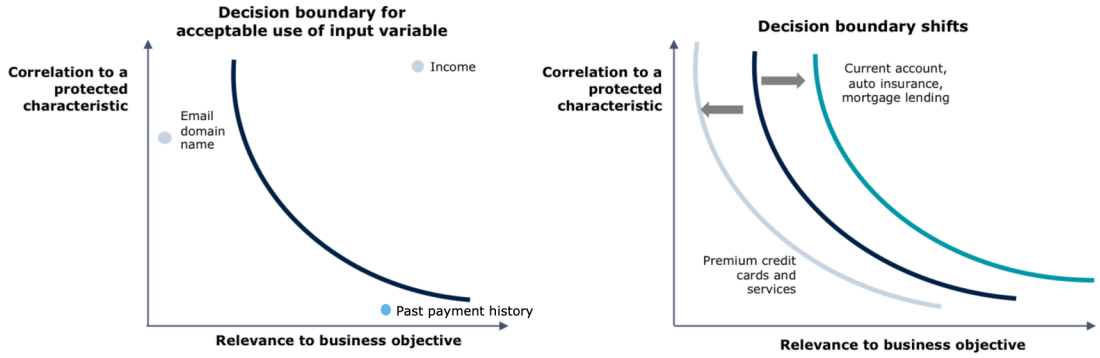


Figure 3.4: Decision boundary for acceptable use of feature

Policymakers should consider this trade-off between accuracy and interpretability to limit what data can be used and what models can be built. For provision of products and services that are considered essential to daily lives, such as current accounts, criminal justice decisions, and car insurance, policymakers may need to ensure that all features included in the variable have a strong inferential relationship with the outcome rather

than simply for predictive correlation. This topic is discussed in important governmental documents, such as the AI Bill of Rights in the U.S. [Hine and Floridi, 2023]. The data minimisation principle in GDPR is aligned to this approach of reducing personal data held by an organisation [Galdon Clavell et al., 2020].

While there is still much work to be done on proxies, they are an important point of discussion because of how they affect the feature selection and feature engineering process in the Build phase. Whether a feature (or a combination of features) is an undesired proxy has such high dependency on the context, domain area, and relevant regulations that it is difficult to generalise its detection mechanism. However, our hope is that our examples and proposal of the decision boundary are steps in the right direction in elucidating the challenge of proxies for developers.

This example section demonstrates that even when the bias is identified, the mitigation strategy is not always apparent, as there is limited consensus in literature. The issue of proxies has been a topic of rising interest in the fair ML community. With this in mind, we return to the proposed questionnaire to discuss its implications, the limitations of the study, and potential for future work.

3.7 Discussion

Our goal was to introduce a risk identification questionnaire to help practitioners identify potential bias risks. The survey shows practitioners find the questionnaire helpful and applicable to their daily practice, particularly in its *breakdown* of bias types introduced in past frameworks, in order to identify where unexpected biases may manifest in ML lifecycle. It provides a targeted and systematic way of understanding the *sources* of bias. Unlike fairness toolkits, it covers the full model development lifecycle. Unlike checklists, it does not attempt to prescribe tasks or activities, but rather directs attention to areas that might warrant consideration based on the context.

3.7.1 Limitations of this study

Sample size and expertise of participants

The questions we asked in the survey would be difficult to contextualise for a respondent with no background nor reference point regarding fairness-related challenges. A lack of background in fairness might have contributed to the drop-out rate and limited the potential sample size, suggesting that while fairness is an area of interest to many practitioners, few have relevant expertise. Indeed, in the demographic question: "Have you ever worked on a product in which fairness and bias assessment would have been useful," 31% answered "no," with several adding in the additional comments that fairness-related concerns are

not applicable to their ML models, e.g. because they do not use any personal data (note later we challenge this view). The survey distribution methodology targeted those with previous interest and experience in ML bias. This, and the high drop-out rate, suggests the respondents that completed the survey are likely more informed and more passionate about these issues than standard industry practitioners.

While this selection bias may affect the generalisability of the findings to wider populations, only practitioners who are building models with concerns about potential discriminatory biases would reasonably use the questionnaire. Therefore, their feedback on the questionnaire is relevant.

(In)correctly perceived relevance: applicability of bias in models without explicit personal information

Despite the 86% who found the questionnaire helpful, several practitioners reported that they did not find the questionnaire helpful because bias detection is allegedly not applicable to their work, as they do not use personal data. Two of the survey respondents also said in their free-text responses that there are no resources allocated on this issue because of limited business incentive and a lack of awareness. Note that such objections were in a relative minority of those who filled out the questionnaire. Only 11.5% of the 105 respondents disagreed that the questionnaire can proactively diagnose unexpected bias issues, and the overall feedback on the questionnaire was positive.

However, it is worth scrutinising the claims that bias is not relevant to some practitioners' work. In fact, models that do not directly use personal data may still raise bias and fairness-related concerns. For example, one of those who claimed it is irrelevant said they use "data sets that do not involve humans (e.g. MNIST)." While handwriting data set may not have personally identifiable data, e.g. associated name, it is plausible that a model built on handwriting data sets such as MNIST could be biased. In fact, researchers could correctly predict the writer's nationality through his/her handwriting [Nag et al., 2018], implying personal information could be deduced from such data. This shows some practitioners may have a narrow understanding of the types of models that could be affected by unintended bias concerns. Our discussion of proxies aimed to illustrate the challenges of data sets that appear unproblematic on the surface but contain correlations that encode sensitive personal information. Understanding the (lack of) awareness around proxies among practitioners is one of the areas we propose for future work.

3.7.2 Future work

Our findings in the risk identification questionnaire paper reveal several opportunities for future research. The first area is a deeper dive into the effectiveness of the questionnaire

through an exploration of how it is used in practice. While our paper validated the questionnaire through a brief survey, in-depth interviews could further uncover findings about the strengths and limitations of the questionnaire.

Another area for future work is in the contextualisation of the questionnaire. The risk identification questionnaire aimed to address the current gap: a lack of a practical tool that operationalises the recent frameworks in bias types. The questionnaire is not intended to prescribe a comprehensive coverage of all potential biases. It should be adapted and extended to be customised to the use case and domain area. This was echoed by a few practitioners, who asked for "more examples" and "more concrete language," stating that "It would be easier to use if it were built with domain-specific examples and language, but that can be adapted." These results show we need more guidance on targeted risk identification methodologies for each domain area. Future work should identify the potential bias sources across use cases and tensions between ethical objectives.

We also reported on trends in practitioner responses regarding barriers to adopting methods for ML bias risk. This included a lack of incentives for business leaders in allocating resources to bias-related initiatives. The survey garnered 105 answers in a month (over the new year period); despite the high drop-out rate of those who abandoned the survey after starting, the high uptake signals practitioner interest in ML bias issues. However, the practitioners' narrow understanding of model biases (especially those related to proxies) and their pushback against a qualitative exercise are especially concerning to the researchers advocating for fairness testing to be more than a routine, box-ticking exercise. Future work could address how to raise awareness of bias risks among practitioners, drive organisations to be proactive in their mitigation, and facilitate integration of risk management methods into their processes. Another opportunity for future work is to expand on our findings with a larger sample size and to address the practitioners' expressed needs and preferences.

In particular, there is a strong desire for guidance on technical and non-technical strategies to mitigate the risks of unintended biases. The questionnaire's scope of breaking down bias types was found helpful in identifying next steps, but it prompted some free-text comments to demand more guidance on what technical analysis and fix are needed. Whereas analysis and mitigation are out of scope for risk identification, this presents an important challenge for future work, in particular because not all mitigation strategies are obvious. Suresh and Guttag (2021) suggest that their bias framework should help future work to "state upfront which particular bias they are addressing, making it immediately clear what problem they are addressing" [Suresh and Guttag, 2021]. Our questionnaire extends their work on bias types into a practical tool, facilitating the process of their identification. It is our hope that the questionnaire similarly helps the discovery of existing gaps in literature – i.e. which questions still cannot be answered – on how to mitigate the

risks of unintended biases in this evolving space.

3.8 Key chapter takeaways

In this chapter, we proposed a risk identification methodology for potential unintended biases in ML development lifecycle, aligned to a standard enterprise risk management framework. We built a questionnaire and walked through a real-life use case on potential biases in an ML algorithm to predict fraudulent insurance claims. We also validated the questionnaire with industry practitioners, which had a strong positive reception overall. In particular, 86% of the practitioners agreed that the questionnaire would be helpful in their "ability to proactively diagnose unexpected issues."

To ensure the end-to-end risk management of ML models and their potential to perpetuate unintended harmful biases, a targeted and systematic bias risk identification methodology is necessary. To promote adoption, risk identification methods should be easy to integrate into an organisation's existing processes and risk frameworks, and allow for the appropriate mitigation strategies to be formulated. The questionnaire's primary role is to identify the potential source of the bias and diagnose the problematic phase in the ML development lifecycle. Our proposed questionnaire introduced an indicative example of such a risk identification method, operationalising the latest framework on unintended biases. The practitioners surveyed were generally in agreement that the questionnaire met their requirements.

This represents but one step – effective risk identification lays the foundation for a more targeted risk analysis and mitigation, and we hope this questionnaire will help practitioners and researchers in this endeavour. Our work reveals important opportunities to explore adaptations of such a questionnaire for different use cases and address any gaps in literature where there is no consensus on strategies to manage bias risk in ML models. Our next two chapters discuss the other phases of the end-to-end lifecycle, including impact assessments and mitigation, to ensure the proposed methods are well-aligned to industry standards and easy to integrate into existing practices.

Chapter 4

ML Test: fairness toolkits and trade-off analyses to select the model for deployment

Introduction

For fairness considerations, ML building and testing are an iterative process, as with each build, the model should be tested against pre-defined performance, risk, and other metrics, including Key Ethics Indicators (See §2.5.5). The end goal of the testing phase is to determine whether a set of performance metrics are within acceptable bounds, which model best aligns to these metrics, and decide *whether or not to deploy the model*. This chapter addresses the analysis techniques and mitigation strategies for unintended biases that would have been identified in the questionnaire we proposed in Chapter 3.

The first section in this chapter is the assessment of the analysis tools available for fairness. **Fairness toolkits** have been recently introduced to improve the adoption of fairness testing methods introduced in academic literature. They work by accepting the training data set and/or model as inputs – sometimes through a user interface – and returning the fairness test results and associated visualisations. They are important efforts in ensuring the fairness testing methods are accessible to practitioners, who are often constrained in time and resources. However, these fairness toolkits reflect the diverse and fragmented landscape of academic literature on fairness. There are significant gaps that limit their applicability to real-world use cases, which we discover through a mixed-methods study of industry practitioners.

The second section relates to potential mitigation strategies for fairness issues. In §2.5.4, we alluded to a critique of technical "debiasing" methods, which incorrectly assumes that all unwanted biases can be quantified and removed. We expand on this with related work and argue that biases introduced in the ML development lifecycle should sometimes

be mitigated through *non-technical* measures. We map the biases in §3.2 to recommended mitigation strategies using the same example case study of insurance fraud prediction.

Overall, this section aims to show the shortcomings of technical toolkits in both identifying fairness issues in ML systems and mitigating them. Due to the inherent challenges of using a technical, generalised method to test for a complex, contextual notion of fairness, it is important in the testing phase to revisit the key ethical and practical objectives of the algorithm defined as KEIs (§2.5) to ensure the model sufficiently meets them. We return to our bias identification questionnaire in Chapter 3 to demonstrate through a case study how it helps design mitigation strategies targeting the bias at its source in the lifecycle. In all, we reiterate the importance of resolving a complex socio-technical problem of fairness – not only through an algorithm – but through people-based and process-based solutions.

4.1 Landscape and gaps in open source fairness toolkits

Due to the growing demand for technical methods that can test for potential unfairness, there has been a recent proliferation of toolkits and packages for assessing fairness in algorithms, particularly where less-interpretable ML methods are used. Substantial academic literature on algorithmic fairness has concerned the development of mathematical and computational definitions of fairness (e.g. [Dwork et al., 2012; Hardt et al., 2016; Kusner et al., 2017]), prompting work to explain the distinctions between them [Verma and Rubin, 2018; Narayanan, 2018] and the trade-offs [Kleinberg et al., 2016] given some of these definitions are impossible to simultaneously meet. In turn, this has led to the introduction of automated bias mitigation techniques, including pre-processing methods to remove estimated bias from the data set [Feldman et al., 2015; Kamiran et al., 2012; Calmon et al., 2017], in-processing methods to train a model to both maximise accuracy and increase fairness [Zhang et al., 2018; Kamishima et al., 2012], and post-processing methods to adapt the predictions after the model build to equalise a metric of fairness between groups [Kamiran et al., 2012; Pleiss et al., 2017; Hardt et al., 2016].

4.1.1 Motivation: lack of guidance on toolkit selection

Recently, various open source 'fairness toolkits' have been developed – both by private companies and through community open source development – to make these fairness methods more widely accessible to model developers. Fairness toolkits enable the developers to upload their own training data set and/or model as inputs, some of the tools through a user interface, to obtain fairness test results and associated visualisations. An overview

of the landscape and key features is in §4.1.5. As these toolkits are to be integrated into developers' model build process, they have the potential to help improve fairness testing and mitigation at-scale across domains (if and where appropriate). On the other hand, there is a risk of these toolkits being applied to an inappropriate use case, misinterpreted without considering the assumptions or limitations of the implemented methods, and/or misused (deliberately or otherwise) as a flawed certification of an algorithm's fairness. In particular, an open source toolkit built by a private company for a specific use case may not be widely applicable to other companies with different priorities or to other industries. Our work addressed the gap in highlighting how certain toolkits are relevant for particular purposes and audience, through a comparative analysis of toolkits' features [**Lee** and Singh, 2021a]. A lack of guidance comparing the available toolkits limits the accessibility and usability of the toolkits and results in a risk that a practitioner would select a sub-optimal or inappropriate tool for their use case, or simply use the first one found without being conscious of the approach they are selecting over others. If so, models with unfair outcomes may be deployed at-scale with the false confidence provided by a tool that was not fit for the use case.

4.1.2 Related work: past studies on fairness challenges in practice

There are two key prior studies to our own on high-level fairness challenges faced by practitioners. A past interview and survey of ML practitioners identified challenges they face in algorithmic fairness that they felt unresolved [Holstein et al., 2019]. Another study assessed the needs in high-stakes public sector decision–making specifically with exploratory interviews [Veale et al., 2018]. Their studies, conducted between mid-2016 to mid-2018, largely pre-date the release of open source fairness toolkits, and therefore such work only investigates the challenges practitioners face, not whether and to what extent existing tools are or could be useful. Our focus group, interviews, and surveys have reaffirmed some of the top-level themes in Holstein et al. [Holstein et al., 2019].

As can be seen in Table 4.1, past studies have studied practitioner impressions generally in tackling fairness-related issues in their role, not specific to any tools or resources. Therefore, there are toolkit-specific findings that are novel, including limitations of the user interfaces and challenges of integration of the toolkit into existing model pipelines. In addition, our bottom-up approach across the open source toolkit landscape allowed us to generalise their shared characteristics and their implications in practice, beyond what has been revealed in past interview studies. For our findings that echo those of past work, we reveal novel insights into how the new open source toolkits still fail to address these key practitioner requirements.

Finding in Holstein et al. (2019)	Finding in Veale et al. (2018)	**Similar finding in our paper [Lee and Singh, 2021a]**
Need for tools and processes to guide fairness-aware data collection, biases in the humans in the loop through labelling and feedback	Challenges in detecting data changes	Limited coverage of the model pipeline
Challenges due to blind spots and unintended biases in use cases	Challenges of scaling up context-specific assumptions	Limited adaptability of existing toolkits to a customised use case
Lack of auditing guidelines	N/A	Steep learning curve required to use the toolkits and limited guidance on metric selection
Needs for more holistic evaluation and communication of the real-world implications	Challenge in getting individual and organisational buy-in	Need for "translation" for a non-technical audience
Addressing detected issues: need for a mitigation strategy	Developer as a "single point of failure"	Limited information on possible mitigation strategy
N/A	N/A	Lack of a tailored user experience that avoids both information overload and oversimplification
N/A	N/A	Accessibility of toolkit search process
N/A	N/A	Challenges in integrating the toolkit into an existing model pipeline

Table 4.1: Holstein et al. (2019) and Veale et al. (2018) key themes, linked to our own findings in our paper

Given the fast-moving nature of ML fairness research, while our study provides a snapshot of the open source toolkit landscape, many of our gaps are generalisable to any toolkit aiming to automate testing of fairness considerations. In our Discussion (§4.1.10), we argue that these challenges are inherent in the attempts to create generalised, simple toolkits for contextual, complex fairness issues. Therefore, while these gaps may be partially addressed, they may never be completely closed by the introduction of new toolkits.

However, our study's focus is on open source and widely-available tools (thereby providing new takeaways), while ML practitioners may also use commercial tools or those built in-house. Generally, open source toolkits can have broad impact, enabling wider adoption and access due to lower costs, accelerated experimentation and delivery, and reduced dependency on third parties [Walli et al., 2005]. Indeed, many enterprises almost entirely rely on open source technologies, as the proportion of proprietary technologies in firms' software portfolios has steadily declined over time [Wurster et al., 2011]. Commercial

tools are also challenging to assess and compare due to their proprietary nature and limited accessibility, making a gap analysis infeasible. We also explore open-source-specific considerations in procurement, requirements, and criteria, e.g. license review, frequency of updates, robustness, etc., which were not explored in previous work but we find to be important to the practitioners. Given the studies by Holstein et al. and Veale et al. predate the introduction of many of these open source toolkits, without an understanding of the open source toolkit landscape, one cannot conclude whether the challenges they reported continue to be truly unaddressed gaps. We conduct a bottom-up review of open source toolkits by examining the code, documentation, and visualisation, in order to understand their key features and inform the following practitioner studies to capture their perspectives on whether the toolkits are fit for their purpose. In contrast to previous work, this chapter aims to provide a feature summary of existing toolkits and discusses the gaps between their offering and the practitioners' requirements, specifically in open source toolkits that are best placed to have a widespread impact due to their accessibility.

4.1.3 Our contribution on fairness toolkits

Derived from our paper [**Lee** and Singh, 2021a], one of the core contributions of this chapter is the *identification of gaps between the capabilities of existing open source fairness toolkits and the requirements of practitioners*, highlighting the implications of these gaps *to inform the development of fairness-related tooling*. This is to move forward the effort to make fairness assessment accessible beyond academia and across industry by engaging practitioners to uncover their needs. Specifically, we assess the relative importance for practitioners of functionality and usability features in the context of an 'ideal' fairness toolkit. We compare and evaluate an indicative sample of existing toolkits to summarise relevant criteria impacting toolkit selection. Finally, we identify gaps requiring urgent attention in order for toolkits to be useful for practitioners in addressing fairness issues in their real-world scenarios. Toolkits suitably designed for users (practitioners) helps better support proper use of the toolkits and accelerates their adoption. Poorly implemented or poorly explained toolkits risk engendering false confidence in flawed methodologies.

In this fast-moving area of fairness toolkits, our analysis provided a snapshot of the landscape and gaps at a point in time. However, while new toolkits and new features may be introduced, many of our findings are generally applicable because they represent the challenges inherent in automating the testing of fairness considerations, such as the difficulty in designing a simple, generalised, technical solution to a complex, contextual, socio-technical problem. While some of these gaps may be addressed in future toolkit development, they would not be completely closed. We present our gaps and the practitioners' requirements as guidance for future fairness toolkit developers, beyond the six toolkits studied in our paper.

4.1.4 Methodology

In order to identify these gaps, we use a mixed method study: an exploratory focus group, a semi-structured interview, and a survey. Our methodology was designed to identify high-level issues with practitioners with prior knowledge of algorithmic fairness challenges in their products, drilling down into the industry needs and their perceived gaps in the fairness toolkits. We validate the findings from prior stages in each subsequent method.

Our overall methodology entailed four steps:

1. **Exploratory focus group** to identify prominent fairness toolkits and derive initial insights;

2. **Comparative review of the selected toolkits** that we conducted to compare the features available in each toolkit to inform the subsequent interviews and survey questions to examine if these toolkits are fit-for-purpose for real-world use cases;

3. **Semi-structured interviews** of practitioners with prior experience in fairness challenges to understand the features in an ideal toolkit and rate how well each of the six toolkits meet their needs; and

4. **Survey** to validate the findings with a broader group and probe into a few insights from previous stages.

5. **Follow-up interview study** conducted a year after the publication of this paper with researchers in human-computer interaction at Carnegie Mellon University [Deng et al., 2022]. Although this was not a part of our original paper, we will incorporate learnings from this paper, as it echoed similar findings.

Our approach is structured to first derive the context and scope through the exploratory focus group and then undertake a deep-dive into the initial findings with the interviews and surveys. To get an overview of the fairness toolkits available and assess their capabilities, we started with an exploratory focus group of practitioners with an interest in the intersection between data science / ML and ethics. Then, we organised semi-structured interviews with industry practitioners with first-hand experience in fairness challenges. We had 15 interviewees, each from a different company. First, known practitioners in ML fairness were contacted for an interview, and others were recruited through snowball sampling (i.e. through referrals from the interviewees) [Parker et al., 2019]. As per the consent form, the respondents and their companies will not be identified or linked to any demographic information to preserve their confidentiality, but we will refer to each of them by their ID (i.e. I1-I15), listed in Table 4.2. The interviewees worked in technology (4), professional services (2), retail banking (2 - I4, I15), insurance (2), financial technology (1), marketing (1), media (1), public sector (1), and academia (1). All interviewees were based in the U.K.

except one from Canada and one from the U.S. All interviewees completed the structured questionnaire except one (I8), in which we were out of time after covering 4 of the 6 toolkits (all but two toolkits: Fairlearn and audit-ai).

Permission was obtained to record the interviews for transcription (with identifications removed), with the recording subsequently deleted. Interviews lasted between 60 and 90 minutes and conducted over a video conference call application. We asked whether they have faced fairness-related challenges in their past products or models and proceeded when they confirmed this as a confirmation of their eligibility for the interview. We transcribed and tagged key themes in the interviews through affinity diagramming, generating codes for topics and grouping them into themes.

Given the time constraints of the interview (60-90min), it was impractical to get each interviewee to comprehensively assess each toolkit by reading all relevant documentation and trying it out on a different data set. Therefore, their views may be limited by what was presented as snapshots of the tools. A few interviewees (I1, I3, I4, I10, and I14) had prior experience with a subset of the toolkits, which allowed them to comment more extensively on their features. Despite this, we saw that the interviewees were able to provide valuable and consistent insights into the importance of each feature in their ideal toolkit and how well each tool appears to meet their needs.

Finally, we designed a survey to reach a wider audience with more diverse levels of familiarity with algorithmic fairness to validate the earlier findings. Of the 71 people who started the survey, 41 (57.7%) of the respondents completed at least one section and 26 (36.6%) completed the entire survey. The study went through our Departmental ethical review process. There were 6 academics, 3 business or product leads, 11 data scientists, 2 ML engineers, and 1 technical lead. Most (14) were from the UK, with 2 each from the US, Germany, and Canada, and 1 each from Belgium, France, and Singapore.

Immediately following the publication of this paper, we were contacted with an opportunity to collaborate on a follow-up piece of work [Deng et al., 2022] to deep-dive into the practitioners' usage of fairness toolkits. In this study, we engaged practitioners directly in using a fairness toolkit on a hands-on ML model testing, conducting "think aloud" interview studies, in which participants are asked to continuously articulate their thinking while exploring and using a software toolkit. Some of these findings are included in this chapter.

Further information on the study and detailed findings are in [**Lee** and Singh, 2021a]. In this chapter, we highlight some of the key observations from the study. §4.1.5 includes findings from 2) comparative reviews of the selected toolkits. §4.1.6 aggregates the themes from our findings across the 1) focus group, 3) semi-structured interviews, 4) survey, and 5) follow-up "think aloud" interview study. Although the focus group preceded the toolkit review in sequence, it informed not only the selection of toolkits but also the initial insights

ID	Role	Domain area	Country
I1	ML Engineer	Professional services	UK
I2	Data scientist	FinTech	UK
I3	Data scientist	Technology	UK
I4	Data scientist	Financial services: retail banking	UK
I5	Researcher	Technology	UK
I6	Researcher	Technology	UK
I7	AI Ethics Lead	Media	UK
I8	Data scientist	Technology	UK
I9	Academic	Academia	UK
I10	Data scientist	Financial services: insurance	UK
I11	Data scientist	Public sector	UK
I12	Consultant	Professional services	UK
I13	Data scientist	Financial services: insurance	Canada
I14	Business or product lead	Marketing	UK
I15	Data scientist	Financial services: retail banking	US

Table 4.2: Interviewees and their demographics

on toolkit landscape and gaps that were validated through interviews and the survey. Therefore, we first discuss the toolkit features from the review and group together the key findings from the focus group, survey, and interviews.

4.1.5 Open source fairness toolkit feature comparison

We conducted comparative reviews of the current capabilities of six prominent open source fairness toolkits, identified as a part of the initial exploratory focus group: scikit-fairness / scikit-lego [Vincent and ManyOthers, 2019], IBM Fairness 360 [Bellamy et al., 2019], Aequitas Tool [Saleiro et al., 2018a], Google What-if tool [Wexler et al., 2019], and Fairlearn [Microsoft and contributors, 2019]. Their key features are compiled into a comparison table, displayed in Figure 4.1. This exercise is to provide a systematic comparison across a range of toolkits, both to help give detail of the different considerations relevant to fairness tools and conduct a critical analysis of toolkits to understand whether they are fit for purpose for real-world challenges of practitioners. By highlighting and assessing particular characteristics of toolkits, we present generalisable findings about the current landscape. For future iterations of fairness tooling, the key themes on the aspects and features that are important to the practitioner will guide their evolution. The differences in the toolkits' approaches will be discussed to point out the potential issues and considerations from a practitioner's point of view, which is later validated in the interview and survey. This section will report on the key feature differences among the six toolkits, informed by the focus group discussions on what criteria a practitioner seeks in a fairness toolkit. This exercise forms the basis for the interviews and surveys, which will validate the importance of these features.

Tool	Setup	Open source user license	Release date	Organization	Open for anyone to contribute code?	Models covered				Group fairness						Individual		Other fairness metrics	Bias mitigation
						Regression	Classification (binaryoutcome)	Multi-class outcome	Handles multi-class protected feature?	Demographic parity (statistical parity)	Equal opportunity / True positive parity / false positive error rate balance	Equal odds (True positive and false positive parity)	Disparate impact	Discovery rate	Omission rate	Counterfactual fairness	Sample distortion metrics		
Scikit-fairness / scikit-lego	python (sklearn)	MIT	2019-03-31	N/A	✓	✓	✓	X	X	✓	✓	X	X	X	X	X	X	N/A	Pre-processing: information filter
IBM Fairness 360	python 3.5+, R	Apache 2.0	2018-06-01	IBM	✓	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	✓	Generalized Entropy Index Differential Fairness and Bias Amplification (full list here: https://aif360.readthedocs.io/en/latest/modules/generated/aif360.metrics.ClassificationMetric.html)	Optimized Preprocessing, Disparate Impact Remover, Equalized Odds Post-processing, Reweighing, Reject Option Classification, Prejudice Remover Regularizer, Calibrated Equalized Odds Postprocessing, Learning Fair Representations, Adversarial Debiasing, Meta-Algorithm for Fair Classification, Rich Subgroup Fairness
Aequitas tool	python 3.6+	Custom	2018-02-13	UChicago	✓	X	✓	X	✓	✓	✓	✓	X	✓	✓	X	X	N/A	N/A
Google What-if tool	Tensorboard / Jupyter or Colab notebook	Apache 2.0	2018-09-11	Google	✓	✓	✓	✓	✓	✓	✓	X	X	X	X	✓	X	Group thresholds	Threshold optimization based on fairness constraints
PyMetrics audit-ai	python	MIT	2018-05-18	PyMetrics	X	✓	✓	X	X	X	X	X	✓	X	X	X	X	Statistical tests to determine chance the disparity is due to random chance (ANOVA, 4/5th, fisher, z-test, bayes factor, chi squared sim beta ratio, classifier posterior_probabilities)	N/A
Fairlearn	python	MIT	2018-05-15	Microsoft	✓	✓	✓	X	✓	✓	✓	✓	X	X	X	X	X	Group max / min / summary	Exponentiated Gradient, GridSearch, Threshold Optimizer

Figure 4.1: Open source fairness toolkit feature summary table

We first studied each toolkit to gather relevant information about its functionality. Figure 4.1 contains the list of toolkits and the types of models covered: regression problems, classification problems (binary only or multi-class), and/or problems with multi-class protected features. A subset of the toolkits handle regression (predicting a continuous variable, e.g. income) as well as classification (predicting a discrete variable, e.g. loan approved or denied). Some toolkits can only handle binary protected/sensitive features (e.g. male vs. female), while others support multi-class features (e.g. age or racial groups). As will be discussed in the next section, practitioners search for tools that are compatible with their model, and if working on a regression problem, two of these toolkits can be ruled out immediately. Figure 4.1 also contains the fairness metrics and mitigation techniques supported by the tool. The most comprehensive of them is IBM Fairness 360 with more than 70 metrics, although its focus is on binary classification problems with some multi-class classification support and no support for regression.

Most of these tools are also focused on group-level fairness metrics, while only Google What-if tool has a focus on individual-level fairness. IBM Fairness 360 supports some individual fairness metrics, such as sample distortion (distance computations between the same individual point in the original and transformed data sets for different distances).

The variety of fairness metrics renders it especially challenging for the user to know what metric is appropriate for each use case. We observed that one potential point of confusion is the differences in terminologies and definitions for the same metric. For example, equal opportunity difference is synonymous with false negative rate difference, and equal odds tests for both false positive and false negative rate disparities [Verma and Rubin, 2018]. Given the tools call the same metric by different names, it may give the impression that they offer different testing mechanisms. In addition, the toolkits have different approaches to guiding users on which metrics is appropriate for any use case. This theme will be further explored in the analysis of interview and survey results.

Most of these packages are built for integration with python, with one tool (IBM Fairness 360) with R support. All of them except Google What-if tool is built to allow for analysis on-premise, i.e. without uploading data into an external environment. What-if tool requires data upload, and its website specifies:

> "WIT [What-if tool] uses pre-trained models and runs entirely in the browser. We don't store, collect or share datasets loaded into the What-if tool. If using the tool inside TensorBoard, then access to that TensorBoard instance can be controlled through the authorized_groups command-line flag to TensorBoard. Anyone with access to the TensorBoard instance will be able to see data from the datasets that the instance has permissions to load from disk. If using WIT inside of colab, access to the data is controlled by the colab kernel, outside the scope of WIT [Wexler et al., 2019]."

Similarly, Aequitas tool, while it has a desktop version available, also has a web-based application through which a user can upload a data set, with the caption:

> "Data you upload is used to generate the audit report. While the data is deleted, we host the audit report in perpetuity. If your data is private and sensitive, we encourage you to use the desktop version of the audit tool [Saleiro et al., 2018a]"

The open source licenses in each of the toolkits' Github repository are either MIT or Apache 2.0 with the exception of Aequitas Tool, which has customised its own license. Aequitas Tool license appears broadly permissive, and the key restriction being that the copyright notice must be included in any future adaptations of the code, and UChicago accepts no liability and provides the code without warranty. All of these tool contributors are based primarily in the United States, with one academic organisation and four private entities. Only scikit-fairness is built completely through the open source platform without any corporate sponsorship or involvement. The release date of the toolkits are in 2018 except scikit-fairness (2019).

These feature comparisons that we conducted and compiled were made available to the interviewees along with a select number of screenshots, standardised to include: metric calculation, guidance of metric selection, and visualisation. When showed the table and asked whether it was useful, all of the interviewees said "yes," as it gives a summary of relevant information they would otherwise be searching for about each tool.

In the context of this mixed-methods study, the feature comparison table was a reference point for interviews and surveys and informed their design. In the context of this thesis, this feature comparison table corroborates the diversity of approaches in how fairness is defined and measured in academic literature. It also demonstrates how the implemented methods are primarily aimed at supervised ML, particularly binary classification (See Chapter 5 on the limited literature on fairness in reinforcement learning).

4.1.6 Key findings from focus group, interviews, and surveys

Starting with the high-level insights from the focus group, we uncovered more detailed observations in the interviews. The key themes tagged in the interview transcripts through affinity diagramming were then validated through survey questions. In this mixed-method study, several crucial gaps emerged from the focus group, interviews, and surveys performed in [**Lee** and Singh, 2021a].

1. **User-friendliness**

    - Steep learning curve required to use the toolkits and limited guidance on metric selection;

Finding type	Our finding
User-friendliness	Steep learning curve required to use the toolkits and limited guidance on metric selection Lack of a tailored user experience that avoids both information overload and oversimplification Need for "translation" for a non-technical audience Accessibility of toolkit search process
Toolkit features	Limited coverage of the model pipeline Limited information on possible mitigation strategy
Contextualisation	Limited adaptability of existing toolkits to a customised use case Challenge in integrating the toolkit into an existing model pipeline

Table 4.3: Key findings in our paper "Landscape and gaps in open source fairness toolkits" (2021)

- Lack of a tailored user experience that avoids both information overload and oversimplification,;
- Need for "translation" for a non-technical audience;
- Accessibility of toolkit search process;

2. **Toolkit features**

   - Limited coverage of the model pipeline;
   - Limited information on possible mitigation strategies;

3. **Contextualisation**

   - Limited adaptability of existing toolkits to a customised use case; and
   - Challenges in integrating the toolkit into an existing model pipeline.

We will now elaborate on each of these gaps in turn.

4.1.7 User-friendliness

The toolkits aim to facilitate fairness testing for developers, yet they have significant limitations in user-friendliness. The steep learning curve for the tools, coupled with limited guidance on the fairness metric selection, renders the toolkit difficult to access for practitioners. Our studies showed the presentation of information was divisive among the practitioners, with some preferring more and others wanting less. The practitioners noted the toolkits' lack of support for their need for "translation" of the results into real-world considerations and implications for non-technical stakeholders.

Steep learning curve required to use the toolkits and limited guidance on metric selection

One of the recurring themes in the focus group and interviews was the difficulty in understanding the fairness considerations for practitioners without prior background on the topic. 40% of interviewees rated the "Guidance for users unfamiliar with fairness academic literature" as "Extremely important" with another 13% rating it as "Very important."

Many interviewees commented on the complexity of the fairness-related challenges. I1 claimed to have taken "two months if not more" to learn about fairness testing, noting that one would have to do "at least weeks of reading." I11 said that "the choice of fairness measures as well as their corresponding trade-offs will depend on the context. In certain areas, these will be less clear and even less obvious to practitioners."

In contrast with the perceived importance for guidance, the guidance for users unfamiliar with fairness literature rated an average of 2.87-3.67 out of 5 (from a Likert scale). I3 reviewed the toolkits as not having sufficiently "easy explanations about the concepts and metrics [that are] not in academic language and style." I1, I5, and I6 alluded to the possibility of "fairness gerrymandering" [Kearns et al., 2018; Bietti, 2020] with practitioners selecting the metric based on which metric the model is able to meet among the metrics available. I3 also noted that there is no way to understand if a functionality or metric is "widely accepted." This echoes the finding from a previous study [Holstein et al., 2019] that practitioners struggle with the learning curve, in part due to the lack of guidelines or standard best practices.

Table 4.4 contains the average System Usability Scale (SUS) score out of 100 and its standard deviation. SUS provides a standardised measurement to compare the toolkits to supplement the topic-specific questions, as the toolkits aim at both developers and higher-level practitioners and can inform non-technical stakeholders [Brooke, 1996a]. As mentioned in Chapter 3, a study of 1,000 SUS surveys showed that "poor" average SUS score is 35.7 (sd 12.6), "OK" is 50.9 (sd 13.8), and "good" is 71.4 (sd 11.6) [Bangor et al., 2009]. All the fairness toolkits fall between "OK" and "good." Fairlearn has the highest score, and the IBM Fairness 360 has the lowest score.

In the SUS survey, almost half of the interviewees agreed or strongly agreed with: "I needed to learn a lot of things before I could get going with this [fairness toolkit] system." Given the interviewees were specifically sampled such that they were very familiar with algorithmic fairness issues, this may be underestimating the learning curve required for the wider practitioner population.

Of the survey respondents, 16% classified themselves as "extremely familiar" with the current academic debates in algorithmic fairness, with 32% "very familiar" and 52% "somewhat familiar." No one responded that they were "not at all familiar" with the

Toolkits	**Average SUS**	**StdDev SUS**
Aequitas Tool	61.33	15.78
Fairlearn	65.71	12.99
Google What-if tool	60.33	17.14
IBM Fairness 360	54.50	13.89
PyMetrics Audit AI	58.04	10.29
Scikit-fairness	62.83	17.32
All	*60.43*	*14.84*

Table 4.4: Toolkit System Usability Survey Scores

fairness literature. For those who responded to the free-text field on whether a randomly selected toolkit guidance layout was helpful, they described it as "quite dense," "too much text," and "might be better broken down in a Q&A format."

One guidance material that had an overall strong positive feedback was the Aequitas Tool (Figure 4.2), which contains a decision tree to assist a user in selecting a metric. However, several commented that while the structure was strong, the wording was difficult to understand with no associated definitions or guidance. Some also expressed concern it oversimplified the criteria for selecting a metric. Google What-if tool's associated blog post that features six experts arguing and disagreeing over fairness definitions was also commended for its colloquialism and its easy-to-understand representation of the conflicts between the fairness definitions. Fairlearn's dashboard was also seen as easy to follow due to its step-by-step walk-through of the fairness testing process.

Lack of a tailored user experience that avoids both information overload and oversimplification

In reviewing the visualisation and guidance, interviewees and survey respondents had marked differences in their preference on the level of detail provided, with some preferring more information and others preferring less. A survey respondent notes that "good information design carries people well through the anxiety of overwhelming data." Some respondents found the amount of information provided prohibitively complex, calling it "quite dense" and "a bit of overload." Several commented that the guidance was too long: "for hands-on technical people who are picking this up on a whim and wanting to use it quickly, they want half the text." For What-if tool in particular, the number of options on the screen was overwhelming for some interviewees. One survey respondent said an ideal toolkit should have "an easy and intuitive user interface with transparent and clear back-end code." Another said, "reducing complexity would lead to higher transparency, and that's crucial."

By contrast, others had a strong preference in favour of the detailed interface. One survey respondent said of Fairlearn dashboard that:

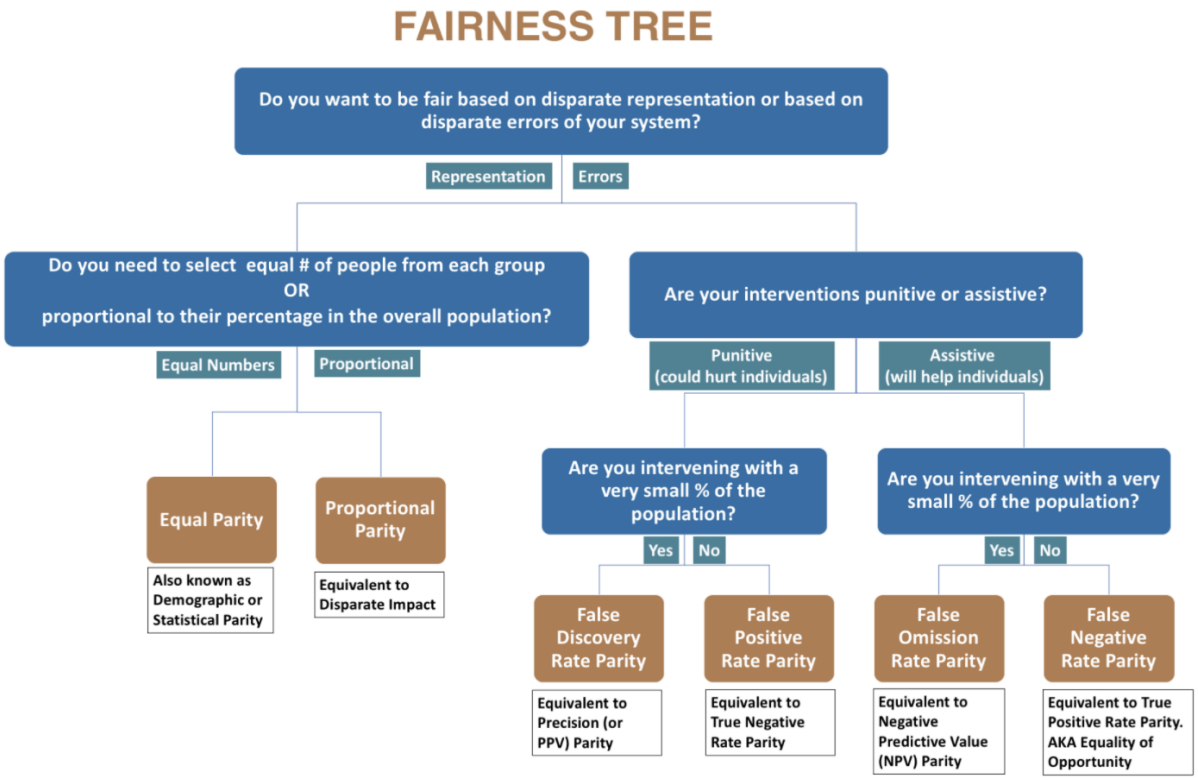


Figure 4.2: Aequitas guidance

> "this makes everything look clear-cut, which it really isn't 'in the wild.' "

. Another interviewee wanted more detail in the guidance for Fairness 360, saying:

> "I feel confident technically using the system, but am I confident in a sense of trusting what it tells me? I don't think so" (I3).

There was resistance to the idea of over-simplifying fairness, which many saw as a complex concept.

Given there may be differences in the level of detail each user requires for his or her purpose, an ideal toolkit should have both (i) a number of options on the user interface that allows the user to deep-dive and slice and dice the analyses, and (ii) an easy-to-use interface that guides the user step-by-step. The former was rated at 3.81/5 on average and at least "Very important" by 69.2% of the survey respondents; the latter had a lower average importance score (3.15/5) but still at least "Very important" by 42.3% of the survey respondents. The interviewees rated a "well-structured user interface" as an average importance of 3.33 on a scale of 5.

Need for "translation" for a non-technical audience

As well as being challenging for data science practitioners with no fairness background, the toolkits were overwhelmingly rated as challenging for a non-technical user, especially

Category	How important are the following in your ideal fairness toolkits?	Mean	Std. deviation
Customisability	1. Ability to adapt to your context-specific use case and data	4.59	0.57
	2. Ability to test for a specific hypothesised or discovered issue	4.21	0.73
	3. Ability to proactively identify unexpected issues	3.96	0.74
	4. Relevant examples and case studies	3.46	1.04
Functionality	1. Comprehensiveness in variety of metrics, i.e. different ways of measuring fairness	4.04	0.88
	2. Documentation	4.15	0.99
	3. Ease of integration into your ML workflow	4.30	0.72
	4. Coverage of different model build pipeline (e.g. identifying bias due to data representation vs. disparity in model prediction errors, etc.)	4.15	0.82
	5. Data security and privacy: whether the data is stored and processed in cloud vs. on premise	3.85	1.22
	6. Reputation of the toolkit provider	3.27	0.83
	7. Identification of action points and next steps	3.31	1.01
	8. Guidance for users unfamiliar with fairness academic literature	3.62	1.20
	9. Comprehensiveness in "de-biasing" (pre-, in-, post-processing) implementations (Note: these are techniques developed to remove the disparity, e.g. between men and women, in the data or the model predictions)	3.85	1.08
User-friendliness	1. Ease of use for technical team	3.92	0.89
	2. Interpretability of results and visualisations to non-technical audience	4.19	0.85
	3. Minimalist user interface that guides the user step by step	3.15	1.16
	4. Number of options on the user interface that allows the user to deep-dive and slice and dice the analyses	3.81	0.94

Table 4.5: Summary statistics: survey results for "How important are the following in your ideal fairness toolkits?" (Likert scale of 0-1)

in producing visualisations, guidance, and user interface that can be navigated by those without a background in math, statistics, and computer science.

It was reported that "for all toolkits, apart from Aequitas, you need somebody technical to run the analysis" and then "translate the findings" to the non-technical stakeholders (I5). Speaking of the various visualisations, an interviewee (I1) said,

> "there's no way a non-technical person could understand this."

This results in a gap between the analysis done by the practitioners and what can be understood by the business function. While a toolkit's main target audience may be the technical developer, over half of the survey respondents 80.8% rated the interpretability of results and visualisations to a non-technical audience as either "Very" or "Extremely important." In rating the visualisations, the survey respondents said they are "likely to be problematic," "more difficult," and "impossible" to understand for a non-technical audience. One survey respondent said the guidance text emphasising the mathematical definitions was "only useful because I have a background in statistics."

In the separate follow-up study, we also found that practitioners desire further support from fairness toolkits to better contextualise ML fairness issues and help communicate often complex fairness analysis to non-technical colleagues in their work places [Deng et al., 2022].

Accessibility of toolkit search process

As suggested in the focus group, we find that almost all interviewees (14 out of 15) claim that they would "use a search engine and iterate through the results until one toolkit that meets their criteria is found, and no further search is conducted." Only one interviewee reported to "comprehensively search for all available toolkits to compare the strengths and weaknesses before selecting the optimal tool." Two interviewees would additionally ask colleagues for advice and search for other work that encountered similar issues. This finding serves to highlight an additional contribution of our work in comparing the features of the six toolkits. The feature comparison chart aims to provide the practitioners with sufficient information about some of the prominent fairness toolkits, as selected through a similar search and discovery process, in order to help them identify the one they need for their use case.

All survey respondents were asked whether they had used any of the six toolkits (multiple selection allowed) and whether there were any other toolkits they were familiar with that were not listed. Only one respondent said they knew of another toolkit: FAT Forensics [Sokol et al., 2019], released in late 2019 resulting from a collaboration between the University of Bristol and Thales Group. However, in attempting to access the user guide in September 2022, the documentation is still empty with the placeholder text

"Coming soon." It appears to still be in active development and not as complete as the other toolkits we studied. Overall, that there were no other toolkits the practitioners were familiar with suggests that our landscape coverage was sufficiently representative for exploring the issues.

4.1.8 Gaps in toolkit features compared to practitioner needs

The toolkits did not address a lot of the practitioners' real-world challenges, due to its narrow focus. Here we discuss the features required by our study participants that are not offered in the toolkits.

Limited coverage of the model pipeline

Echoing the findings in [Holstein et al., 2019], the interviewees emphasised the apparent focus of the toolkits on the model building and evaluation process as compared with the remaining model lifecycle. According to I1,

> "Each section of the model building pipeline is important – testing your training data, representation, model output, proxy variables, etc... no tool has an end-to-end 'this is what is going on in your system.' "

In the survey, 75% respondents answered the "coverage of different model build pipeline" to be at least "Very important." A survey respondent specifically pointed this out as a limitation of toolkits they previously used, saying,

> "most of the toolkits tend to straddle both [auditing / mitigation and data exploration] which presents challenges... so it is not as useful as a part of ML pipeline."

Some gaps specifically mentioned were checking whether the data set is representative of the broader population and whether there were features acting as proxies of protected features, e.g. postcode for race or occupation for gender. I2 claimed a major gap was in the lack of benchmarking data sets or a reference point for whether there is a selection bias in the data collection process. I10 suggested there should be a way to understand which input features are potentially acting as proxies for protected features, "especially when the feature engineering has been done by a human" (I10). A survey respondent also noted,

> "the analysis needs to explore the idea of proxies, something we do manually today."

(See §3.6 for a discussion on **proxies**).

Limited information on possible mitigation strategies

There was mixed amount of enthusiasm for tools that offered "debiasing" pre-processing, in-processing, and post-processing implementations. Several interviewees (I1, I3, I5, I6, I10, I11, I13) were skeptical of these techniques. I3 claimed that these methods are:

> "dangerous because it looks simple but doesn't solve any problem. It's like a gimmick, like training a constrained classifier, but it doesn't solve the underlying issues of bias you may have."

I1 said it "doesn't solve the bias at the root." One interviewee (I10) claimed some of the bias mitigation tools could be inconsistent with anti-discrimination laws, especially any that explicitly use a protected feature (e.g. race) to give preferential treatment, and suggested that the mitigation strategy should depend on the context, which "may not always be a technical solution." On the other hand, several other interviewees (I2, I4, I8, I9, I12) viewed these implementations favourably. I12 noted that some tools' "lack of mitigating action leaves a huge knowledge gap for data scientists to fill."

4.1.9 Contextualisation

While the toolkits present themselves as generally applicable, they were often seen as difficult to adapt to particular use cases. In addition, some of them were seen as difficult to integrate into existing workflows and pipelines, which many practitioners see as a crucial pre-requisite to adopting the toolkits.

Limited adaptability of existing toolkits to a customised use case

The strongest consensus regarding the ideal fairness toolkit was the importance of the "ability to adapt to a context-specific use case and data," with all responses either 5/5 or 4/5 and an average of 4.7. Similarly, in the survey, all except one respondent rated this as "Extremely" or "Very important." The existing toolkits were rated on the same criteria as an average of 3.24 out of 5, with PyMetrics audit-ai scoring the lowest at 2.71 and IBM Fairness 360 the highest at 3.73, with several interviewees noting that additional work would be needed for the toolkits to be applicable to their use cases.

Because audit-ai was built tailored to the U.S. employment guidelines for internal use as an algorithmic hiring company, many found it to be inapplicable to their own use cases. For example, the "4/5ths rule," i.e. the guidance that the lowest-passing group has to be within 4/5ths of the pass rate of the highest-passing group, may not be an appropriate threshold in other domains. However, their unique approach of statistical testing to calculate the likelihood that the disparity is due to random chance was lauded by a few interviewees (I5, I6) and survey respondents as important, compared with other tools

reporting the outcome disparity (e.g. re false positive rate) without confidence intervals or statistical significance.

IBM Fairness 360 was rated highly for having "a lot of useful code"; however, one interviewee (I9) who had extensively used the tool noted that a lot of their tool is "hard-coded to their data and their use case, so it was a matter of how much extra work is needed." I13 critiqued the tools for having relatively little focus on regression problems compared to simple binary regression problems. For his work in insurance pricing, 95% of his work involves regression problems with multi-class protected features, thus several of these toolkits are not applicable; referring back to Figure 4.1, only What-if tool and Fairlearn has coverage of these model types. This issue was also highlighted in the focus group as a major concern on the toolkit's adaptability.

Challenges in integrating the toolkit into an existing model pipeline

Another point of consensus was the importance of the ease of integration of a toolkit into the model building workflow and pipeline. This was rated as "Extremely important" for 60% of interviewees and "Very important" for the remaining 40%. Among the survey respondents, 85% rated it as at least "Very important."

However, the toolkits were rated an average of 3.24 in their ease of integration, with the lowest score at 2.47 for Google What-if tool and the highest score at 3.93 for Scikit-fairness. This is due to the differences in how they are designed to integrate with existing workflows. Google What-if tool visualises the data such that the model developer could explore potential biases, rather than being supplementary to model development; scikit-fairness was built to embed fairness testing and mitigation directly into the model build. The other tools have a mixture of the two, with some stand-alone visualisations and some efforts to provide testing code base that can be integrated into the model.

As discussed briefly in §4.1, several interviewees criticised Google's What-if tool and the Aequitas web application for the requirement to upload the data, noting this would face challenges from their organisations on whether this is GDPR-compliant and in adherence to relevant privacy policies. This partially contributed to the low score of Google's What-if tool, especially given the visualisation required a setup in Tensorboard or Jupyter notebook. One survey respondent said any toolkit with any processing off-premise, even if the data set is not stored,

> "would need a very large amount of governance and security validation to be allowed to be used with corporate data."

This sentiment was repeated for several survey respondents, with many listing any solution that is not completely on the local computer as a 'deal-breaker.'

Several interviewees (I1, I3, I4, I10, I11, I13, I15) claimed that having to upload their data sets, even if it is not stored, could immediate disqualify the tool for usage due to organisation's policies. Only one interviewee (I14) said that this was not an issue because the company has a pre-arranged partnership agreement with Google.

Scikit-fairness received a high score because most of the interviewees were python users; however, two interviewees (I1 and I11) commented that it is only easy to integrate into scikit-learn and gives no flexibility if working with any other package. I11 said for her organisation that often does not use python or R, many of these packages would require an integration layer to work with their existing ML models. However, this seems to be a limited issue among those surveyed. When asked "do you use tools that easily integrate with packages built in python and R," all responded yes.

4.1.10 Discussion: implications and limitations of our study

Different users with different needs

It was clear that a user interface with a one-size-fits-all tailoring toward practitioners with prior understanding of fairness limits the accessibility of these toolkits. Different users have varying preferences and needs from their interface. A key example of this is the high standard deviation in the survey ranking of the importance of mathematical definitions in a toolkit guidance (mean: 4.04/8, standard deviation: 2.79) and the ranking of the importance of visualisations that are helpful for a non-technical audience (mean: 4.48/8, standard deviation: 2.57). As flagged in the interview, some practitioners with a background in statistics may want a detailed mathematical definition, while those looking for a quick proof-of-concept may want a simple user interface for business stakeholders' review. To validate this, there was no strong consensus in the ranking of usefulness in guidance features (average ranking from 4.04 to 5.12) except in two cases: the importance of the explanation of the intuition behind a definition ($\bar{X}$: 2.77, S_x: 1.82) and the non-importance of the relevant legal context ($\bar{X}$: 6.62, S_x: 1.81). One survey respondent explained that the legal context is "better dealt with by a more appropriate (not data) person." Future work could explore how the roles and responsibilities of relevant stakeholders in business (e.g. risk practitioners, lawyers, and business product owner) may be able to interpret and provide relevant context for the toolkit's application. Future work could deep-dive on specific human-computer interaction considerations, e.g. API usability studies [Myers and Stylos, 2016; Zibran, 2008; Acar et al., 2017] considering developers' perspective may be relevant in assessing the specific technical strengths and limitations of some of the toolkits.

Real-world implications: the potential pitfalls of fairness toolkits

The guidance and design of the tool, along with its functionality, could affect the user's interpretation of toolkit outputs, potentially raising the risk that the user could be misled with over-simplified explanations to be overconfident in a model's fairness or confounded by its complexity and pushed to abandon the toolkit. In fact, the follow-up interview observed that most (seven out of nine) of the participants committed to the "fairness through unawareness" [Dwork et al., 2012] pitfall: attempting to mitigate biases in the ML pipeline by simply removing or ignoring the sensitive features like sex or address. P9 in the interview, for example, argued that *"I feel that sex is one of the sensitive [features]. To make the model fair, I'd rather just remove it before training (the model)."* In reality, omitting sensitive features may lead to more disparate outcomes in practice. Furthermore, none of these seven participants considered whether seemingly neutral features might be a proxies for other sensitive attributes. This shows limited awareness on how proxies (discussed in §3.6) can add discriminatory bias to the system.

It is also important to consider whether the toolkits with necessarily reductionist definitions of fairness are appropriate and beneficial from a societal standpoint. Several academics have objected to the "automation" of fairness assessments because these tools fail to consider the socio-technical system, the nuanced philosophical and ethical debates, and the legal context of what it means to be fair [**Lee** et al., 2021; Wachter et al., 2021]. For IBM Fairness 360, in answering whether the tool should be used at all, the guidance warns that the tool applies to limited settings and is intended as a starting point for wider discussion [Bellamy et al., 2019]. The practitioners in the interview and survey were generally positive in their reaction to the notion of a fairness toolkit to help navigate an extremely complex issue, but several expressed concern for "fairness gerrymandering," (or "ethics washing") [Kearns et al., 2018; Bietti, 2020] or selecting the metric based on which ones were satisfied, and for the false confidence the toolkits may give to the model developer based on an incomplete or partial assessment of fairness. Future work could examine in-depth the disclaimers and limitations described for each of the toolkits and whether they align to the academic understanding of suitability of each implemented method.

Limitations of our study

We studied the comparative features of six existing fairness toolkits and identified several key gaps in their capabilities in meeting practitioners' needs. While algorithmic fairness represents a high-profile discussion, and is increasingly an area of concern across industry, among the survey respondents, only 48% considered themselves at least "very familiar" with existing fairness literature. To test one's familiarity with issues of fairness, we asked which two fairness definitions are generally incompatible, and 44% selected the correct

answer: equal odds and positive predictive parity, whose incompatibility was a well-cited example in the U.S. criminal recidivism scoring model and academically proven [Kleinberg et al., 2016]. 36% gave the incorrect answer of equal opportunity and equal odds; equal opportunity is a subset of equal odds, meaning that if equal odds is satisfied, it implies equal opportunity is satisfied [Hardt et al., 2016]. The remaining 20% responded they are not sure. Future work could explore an average practitioners' familiarity with fairness issues in a more representative sample.

The high drop-out rate in the survey (42.3%), i.e. those who start the survey but abandon it after reading the questions on fairness toolkits, also suggests that the prospective respondents may be interested in fairness considerations but do not have the relevant understanding of the topic. While the resulting low sample size limits the external validity of the findings and the ability to conduct more in-depth statistical tests, the key findings persist through the focus group, interviews, and surveys. The methodology, such as reporting of percentages, scales, with references to specific interviewees, is consistent with past human subject research on fairness [Holstein et al., 2019; Veale et al., 2018].

For the focus group and interview, practitioners with expertise in fairness were purposefully recruited and sampled; therefore, the results are only representative of those with pre-existing understanding of the typical fairness challenges. However, the fact that both these stages found gaps and limitations, especially in user-friendliness and interpretability of the toolkits and their guidance, suggests that the learning curve may actually be much steeper for an average practitioner with more limited exposure to fairness metrics. While the nature of these toolkits may evolve over time, our findings on practitioner needs and the high-level perceived gaps would provide important signposts for future development.

Our survey showed no one except for one respondent had used a toolkit not on our list, suggesting that there were no glaring gaps in our landscape assessment. We were not aiming for an exhaustive comparison of all toolkits in existence; rather, the six toolkits indicative of the landscape were reviewed in order to elaborate on general issues and concerns. Given the method of uncovering toolkits was through search engines, an approach confirmed in our study as consistent with what occurs in practice, these six are those that practitioners are likely to come across in their search processes. Therefore, even if there is a toolkit that closes some of these gaps, its limited awareness and accessibility is still an obstacle to its adoption.

Given the evolving nature of open source toolkit landscapes, our analysis is necessarily a snapshot: a view of the landscape and gaps at a point in time. New versions of toolkits may be released with additional features or changes to their user interface, and practitioner needs may change over time. However, many of the issues are broadly applicable to all toolkits aiming to automate testing of fairness considerations. While the steep learning curve required to use the toolkits can be somewhat mitigated through improving guidance

provided, understanding whether a model is fair is an inherently complex and contextual exercise that is challenging to address in a generalised toolkit. Because "translation" for a non-technical audience, adaptation to a customised use case, and the mitigation strategy often require an understanding the context-specific implications (such as KEIs), a generalised toolkit has its limitations on its ability to meet these needs. As such, many of these gaps may be addressed but not completely closed and will remain relevant for future toolkit iterations.

Future of fairness toolkits

Fairness toolkits are a fairly recent phenomenon, with the first release in 2018, and several interviewees were surprised to learn about their availability and diversity. Only 54% survey respondents had used any open source fairness toolkit before, despite our sampling of groups with likely exposure to fairness-related concerns. With the growing attention on issues of fairness, it is important that any fairness toolkits are accessible, usable and fit for purpose. This paper [**Lee** and Singh, 2021a] contributes a gap analysis and the associated findings regarding practitioner needs and the features of available open source fairness toolkits. With a focus group, semi-structured interviews, and surveys, we identified key themes of practitioner requirements that require more attention. In addition, the feature comparison helps address the gap in accessibility of the toolkit search process, as it helps users select which toolkit is suitable to their needs.

This analysis can help inform future tool development in order to bridge the gap between the introduction of methodologies in academia and their applicability in real-life industry contexts. As discussed, some of these gaps may be mitigated but may not completely disappear due to the inherent challenges in providing a generalised, simple toolkit for a contextual, complex exercise. Other gaps may be considered for updates in future releases. For example, in designing how information is presented to the users, the toolkit developers will need to tread a fine line between over-simplification and information overload. Toolkits could address this gap by providing general results in simple, easy-to-understand format with the option to drill down into details for a more tailored user experience. Toolkits' integration layer could also be improved, with guidance provided on how it fits into the overall model pipeline. Further guidance and context-specific case studies could bridge some of the gaps in the toolkits' steep learning curve, their need for non-technical "translation," their limited mitigation strategies, and their contextual adaptability.

The feature summary with relevant characteristics of each of the six selected toolkits can help facilitate a practitioners' toolkit search and evaluation. We have found that the toolkits are diverse in their approaches and do not simply reflect different implementations of the same fairness methodologies. Given that (as our results indicate) many practitioners

look for a tool until they find one that meets their needs, a comparative review of the toolkits would help practitioners understand the toolkits' offerings and aid their selection process. We have shared this table of features on GitHub, along with the other workstream outcomes of the focus group (Ethics DataDive) hosted by DataKind UK, in order to allow for others to comment on and update what is available in the open source landscape. It is our hope that this will become a reference point and a repository of information on the practitioners' guide to various ethics toolkits.

4.1.11 Key takeaways on the landscape and gaps in open source fairness toolkits

Our results suggest that industry practitioners are still struggling with finding a way to identify and mitigate potential unfairness in their models and systems. Only by keeping close to the practitioners' requirements and preferences can the open source developers ensure widespread adoption of their toolkits. The toolkits were developed to encourage model developers to be more cognisant of the potential ethical implications of their algorithms in relation to their impact on societal inequalities. An effective fairness toolkit could foster the culture among practitioners to consider and assess unfair outcomes in their models, while a poorly framed or designed toolkit could engender false confidence in flawed algorithms. In particular, claims of fairness toolkits that they can "solve" unfairness may fuel the drive of practitioners into the "trap" of solutionism, failing to recognise the possibility that the best solution to a problem may not involve technology [Selbst et al., 2019]. Future development of toolkits should remain vigilant to ensure their adoption is aligned to the over-arching goal: to ensure our algorithms reflect our ethical values of non-discrimination and fairness. Our study flagged aspects of fairness toolkits that are important to consider – both for developers and product designers of fairness toolkits and for users (ML model developers) in determining which toolkit best suits their purpose for fairness testing.

4.2 Mitigation strategies: critique of "de-biasing" methods

In our study on open source fairness toolkits, a key gap was the limited information on possible mitigation strategies. This was despite some tools, such as IBM Fairness 360, advertising itself as a "de-biasing" solution. We have alluded to critiques of "de-biasing" methods in §2.5.4 and in §4.1.8. In this section, we offer a fuller view of why a technical approach may not fully resolve complex fairness issues.

There are three classes of technical "de-biasing" methods: **pre-processing**, removing

bias from the data before the algorithm build [Kamiran et al., 2012; Calmon et al., 2017; Zemel et al., 2013; Feldman et al., 2015], **in-processing**, building an algorithm with bias-related constraints [Zhang et al., 2018; Kamishima et al., 2012], and **post-processing**, adjusting the output predictions of an algorithm [Hardt et al., 2016; Kamiran et al., 2012]. In §2.5.4, we observed that these methods presume that all undesirable inequalities and biases in the system are known, can be quantified, and surgically removed from the desirable inequalities, which is often impractical. Past studies have shown that the attempt to "repair" and remove undesired bias is ineffective when the legitimate factors for decisions are correlated with the protected characteristic, e.g. income to race or gender in a lending algorithm [Corbett-Davies and Goel, 2018].

In addition, these methods may end up harming the groups they are intended to protect in the long-term; one study considered the long-term impact of a "fair" learning algorithm and found that giving loans to minority applicants resulted in higher default rates and – over time – lower credit scores among the minority applicants [Liu et al., 2018]. In the presence of a feedback loop, we need to consider the trade-offs being made: in this case, between the potential of harmful defaults due to granting unaffordable loans and equalising loan approval rates between racial groups.

It is worth remembering that in the toolkit study, 7 out of the 15 interviewees vocally expressed skepticism about "de-biasing techniques," (§4.1.8), calling it "dangerous" and "like a gimmick." One interviewee noted that "de-biasing" methods that use the sensitive feature to transform the data, model, or outcome to give preferential treatment to a minority group could be in violation of anti-discrimination laws [**Lee** and Singh, 2021a]. The automation of fairness testing in toolkits may be objectionable based on its limited consideration of the legal context of non-discrimination laws [Wachter et al., 2021]. Toolkits themselves often are accompanied by guidance, including warnings such as: the tool applies to extremely limited settings, fairness is a complex issue, and the toolkit is intended as a starting point for wider discussion [Bellamy et al., 2019; Wexler et al., 2019].

While fairness toolkits are useful in improving accessibility of fairness testing methods and visualisations, in their current form, they must be used with caution. They should not be accepted as the panacea to fairness considerations or a one-stop-shop for fairness testing and mitigation. Rather, as per their own disclaimers, they represent the quick-start foundation for better understanding what fairness issues may exist in the data sets and models. In particular, the "de-biasing" methods may be applicable in extremely limited number of use cases, but it is important to understand their limitations, potential for long-term harm, and alignment with non-discrimination laws.

Often, the solution to these biases is not technical and cannot be solved algorithmically because their sources are in the people and processes in the socio-technical system. Instead of looking for an algorithmic solution, it may be more productive to counteract these

biases at their source. The following case study demonstrates some of these non-technical mitigation strategies.

4.2.1 Case study: Mitigation strategies for biases in insurance fraud model

This section will use a real-world case study to bring to life how the questionnaire assists the practitioner in identifying targeted mitigation strategies. In the questionnaire, Section (A) identifies the context, and each subsequent section addresses one type of bias, facilitating the design of mitigation strategies appropriate for that bias type. We will discuss examples of analyses and mitigation strategies that could follow from the practitioner's self-identified risks through use of the questionnaire. Note while the assessment in §3.4 was done by the practitioner and represent his/her views, this section represents our own response to the issues raised. After the developer identified the bias risks in fraud detection, in the paper [**Lee** and Singh, 2021b], we suggested potential mitigation that would be targeted to each bias.

(B) Design: historical/external bias

Given predictions related to criminal acts are often accused of racial or faith-based biases, practitioners could check model performance against racial and faith groups, if these features are available from the data. If not, it could be possible to check model performance by region, which may be acting as a proxy for race or religion, to assess whether high-minority-group areas are more prone to model errors. Regarding socioeconomic biases, the developer could check model performance by income level while controlling for the ratio of claim amount to income.

(C) Data collection: Representation bias

Data recorded by claim handlers should be assessed for any subconscious bias, e.g. flagging one gender as more suspicious. In particular, if there are any differences in fraud detection correlated to the claimants' language skills, the team may consider staff retraining on subconscious biases or hiring staff who speak other languages. Third party data providers could be asked to provide documentation on their data collection methods and any potential biases. The unknown "true" false negatives could be retroactively identified as the team continually assesses what types of "non-obvious" fraud types may be missed. Given the rarity of fraud (relative to legitimate claims) and its under-representation in the dataset, the developer could consider whether over-sampling or pre-processing methods are appropriate, e.g. SMOTE [Chawla et al., 2002].

(D) Feature engineering: measurement bias

Features developed based on fraud intelligence or histories should be assessed for validity and appropriateness, especially if they are highly correlated to legally protected features (e.g. gender, disability status) or features historically associated with criminality (e.g. race, religion). This is to ensure the subjectively engineered features do not embed any unintended biases as proxies of demographic characteristics. Geographical patterns of fraud should also be checked for unintended correlations with racial or religious groups. The model could be trained on confirmed instances of fraud and on investigation results in addition to those correctly flagged.

(E) Model build and training: aggregation bias

The model may be improperly aggregating together different types of fraud with different causal mechanisms. One may consider whether separate models should be built for fraud types that are sufficiently different, rather than representing them in a single model.

(F) Model evaluation: evaluation bias

The relative importance of False Positive/Negative results should be weighted differently by business function. In evaluating model performance, it is important the model is not over-fitting to a particular metric, and to find diverse metrics that closely reflect and measure the organisation's practical and ethical objectives and their relative prioritisation. This may include the risk of unintended discrimination, e.g. against a racial group.

(G) Model productionisation and monitoring: deployment bias

The human feedback mechanism for any errors should be reviewed, especially whether the feedback loop may be reinforcing any existing biases, e.g. whether certain types of fraud are being confirmed or overlooked. The fraud investigators may be prone to confirmation bias if inclined to trust the model's classification of a claim. The system should be robust to any external changes, e.g. change in policy or input data distribution. While this is currently tracked manually, the developer may consider automated monitoring systems, testing procedures, and controls to assess changes in key metrics in live environments. Overall, the investigators and the model should all be frequently retrained for any new or previously overlooked types of fraud.

This section aimed to demonstrate that the bias identification questionnaire introduced in Chapter 3 can be used in the Testing phase to design mitigation strategies. We show that some of the gaps identified in our study of fairness toolkits can be met through the usage of our questionnaire, a non-technical tool.

4.3 Revisiting the Key Ethics Indicators (KEIs) defined in the design phase

In assessing the limitations of the open source fairness toolkits, we found a major gap to be need for "translation" for a non-technical audience, given the fairness results do not easily guide the user to the potential implications in a real-life setting. As shown in §4.2, identifying the potential types of biases facilitates an understanding of what types of analyses (e.g. bias quantification) and mitigation strategies are required. In this way, targeted risk identification enables a more effective management of model bias risks. However, beyond unintended biases, it is important to understand whether a model is ready for deployment into live environment, through an assessment of whether it best meets the requirements out of all available options.

The KEI approach introduced in Chapter 2 was designed as an end-to-end process, from the definition of "success" metrics to model selection. Figure 4.3 is provided below as a reminder of the six proposed stages. In this section, we address the two final stages in the KEI process: the calculation of trade-offs and the selection of a model. We now present an indicative set of action points following the risk identification that demonstrates the potential for this approach.

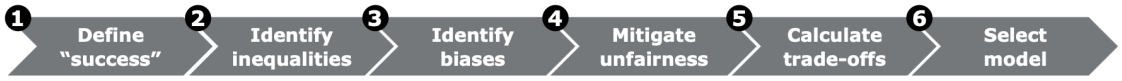


Figure 4.3: Proposed KEI process

Section A of the questionnaire contextualises the use case-specific objectives in relation to the potential impact of accuracy and of bias, which facilitates the impact assessment. Positive impacts include reduced claim costs, reduced funds available to criminal groups, and the quicker processing of genuine claims. These could be formulated as: estimated claim cost per model, amount of truly fraudulent claims withheld from suspected criminals, and average claim processing time. The negative impacts include false persecution of honest claimants and reinforcing criminality biases of certain income, religious, or racial groups, which could be formulated as the percentage of false positives of previously marginalised sub-groups. It is important to explicitly state and quantify such objectives. In work on U.S. mortgage data, Lee and Floridi (2019) visualised the trade-off between aggregate financial inclusion (available credit) and exclusion of historically marginalised minorities (denial rates of black applicants), demonstrating that such analysis can help the decision-maker select a model depending on objective prioritisation.

In the case of fraud detection, Fig. 4.4 shows hypothetical models A-G and their trade-off between false positive rates for minority religious groups (%) and truly fraudulent claims flagged by the model (GBP). While based on hypothetical (non-existent) models, it

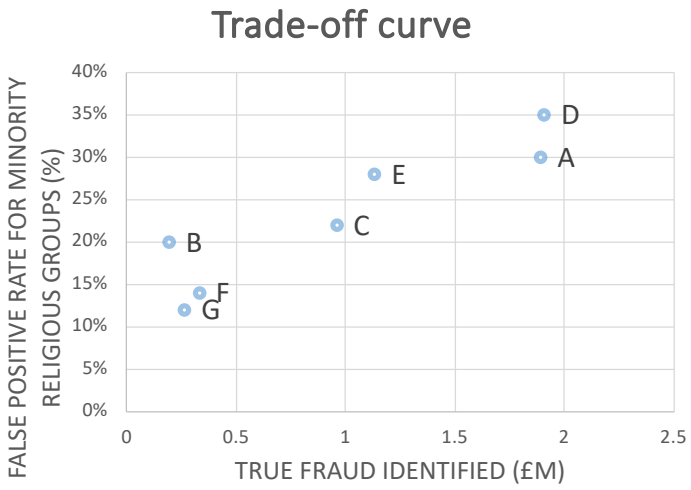


Figure 4.4: Illustrative example: trade-offs in fraud detection model

shows the potential for an informative impact assessment related to unintended biases. We used this as an illustrative example to discuss with the developer from our interview. This is to show that the KEIs are applicable to other domain areas outside of retail lenders, which was discussed in Chapter 2. For example, Model D is the most accurate at identifying true fraud, but it also has one of the highest false positive rates (FPR) for minority racial group – having a model with 35% FPR may be considered unacceptable. Model A performs similarly for identified true fraud but with only 30% FPR and may be chosen over Model D. Model B is worse in absolute terms than F or G so can be removed from consideration. The developer found this type of a model selection process to be insightful, as it translates a vague concept of a fairness-accuracy trade-off into real-world implications. Indeed, this is the main purpose of the KEI approach.

The questionnaire was designed to detect bias sources so as to design an appropriate mitigation strategy. While mitigation processes do not fall within the questionnaire's scope, by proposing a methodology for targeted risk identification, we aimed to provide practitioners with actionable insights for their decision-making on whether the model they built is compatible with their value priorities and risk appetite.

In our study of open source fairness toolkits, many practitioners rightly challenged the notion of a technical "de-biasing" algorithm that can solve their fairness issues. In this section, we demonstrated through a case study that a holistic mitigation strategy for fairness in ML should address the people and processes, as well as the model and data. Effective bias identification in the model build phase (§3) is an important foundation to enable targeted mitigation strategy in the testing phase.

4.4 Key chapter takeaways

In this chapter, we showed through a mixed-method study that the current open source fairness toolkit landscape has major gaps. Instead of relying on these toolkits to get a superficial pass/fail result, developers need to build their own metrics specific to their use case and context. In Chapter 2, we proposed building multiple models and calculating KPIs/KEIs to ensure all metrics are within acceptable bounds, with measures of success that are more specific to the context than accuracy and fairness. In Chapter 3, we introduced the bias identification questionnaire. We recommend these methods are used in an integrated, holistic governance of fairness-related risks that covers the end-to-end pipeline.

When KPI/KEIs are signed off, and the model is selected that best represents the decision-maker's values and risk appetite, then the model can go into production. In the next chapter, we address how we deal with dynamic systems in re-training cycles of "fair" ML.

Chapter 5

ML Monitor: risk factors, reviewability, and fairness under uncertainty

Introduction

Once the model has been built and deployed, the objective of the monitoring phase is to ensure it is working as expected in live environments, operating within the acceptable boundaries of various metrics (e.g. KEIs in §2.5.5). Due to the retraining cycles, a deployed ML model requires monitoring and record-keeping that is proportional to the model risk and at a level of detail that is appropriate to the context that enables any errors to be traced back to their sources. In an *online, live* learning setting, as a model is re-trained on new data, these cycles lead to dynamic risks, compared to more static risks in pre-built, offline systems, such as robotic process automation that follows a set of rules. Re-training on incoming data could introduce new representation bias that was not present in previous training data. In this phase, if the re-training cycles do not require technical interventions, the system may be primarily used by the business process owner, rather than being in the hands of the main technical developer [Lee et al., 2020]. For example, a fraud detection model would be used by fraud investigators, who may not have a technical background. Therefore, it is important to have monitoring mechanisms in place that can prompt relevant personnel that an intervention is necessary, ideally triggering a business process that may involve a model review by the technical developer if adjustments to the model are needed. Sudden changes to the input data distribution or a drop in KEIs may prompt the fraud investigator to contact the model developer. In addition, in an enterprise risk process, there may be internal and/or external reviewers of the model. Internal reviewers may include the model validation team and the internal audit team, and external reviewers may include consultants, regulators, and external auditors.

In this chapter, we focus on *reviewability*, a concept introduced in our paper [Cobbe et al., 2021]. Reviewability is a record-keeping framework for both technical logging and organisational processes. Reviewability should be targeted to ensuring the decisions can be traced back to their sources (the people, process, and any models), especially in case of any errors. This is associated with **robustness**, which is "the degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions" [rob, 1990]. Generally, robustness is discussed in the context of accuracy metrics when there are drifts in the environment or incoming data sets. However, in this thesis we view robustness beyond its functional performance and in relation to the system's Key Ethics Indicators, especially fairness. In other words, reviewability looks beyond whether an ML system retains its accuracy metrics in the presence of external changes; it aims for a holistic view of whether the ML system has the expected real-world impact in meeting its practical and ethical objectives.

In this chapter, we first identify the risk factors in AI systems that may require different intensity and layers of monitoring. High-risk systems may require closer monitoring than low-risk systems. Then, we outline the logging and reporting requirements for AI systems in the context of "reviewability." Finally, we discuss how to consider fairness in a system with high levels of uncertainty. As the main contribution of this chapter, §5.3 Fairness under uncertainty proposes a taxonomy of six layers of uncertainty in an ML system and how they may affect its fairness considerations. We formalise a theory of uncertainty and provide a pseudocode of its implementation. Overall, this chapter is concerned with how to ensure a deployed ML model continues to be fair, through monitoring mechanisms proportional to its risk factors, through reviewability, and through accounting for uncertainty in sequential decision-making.

5.1 Risk factors in AI and automated decision-making

Monitoring requirements for AI vary depending on the level of risk in the AI system. In §3.1, we argued that guidance documents' focus on AI may mislead the readers that the risk lies in the ML techniques. We then proposed that the risk management process should be tailored to the risk level and potential impact of each use case. It is the general consensus among risk management frameworks that monitoring and control mechanisms required for each system should be proportional to its risk [Lee et al., 2020]. In a post-go-live, online setting, the level of oversight needed varies depending on the context, process, and technology of each system.

In our paper [**Lee** et al., 2022], we present an illustrative set of relevant risk considerations, across a range of dimensions. Undertaking a nuanced, holistic risk-based analysis helps practitioners understand the types of mitigation strategies as appropriate

for their system. This is more useful than one driven by a solely by a broad classification as to whether their employs 'AI,' as discussed in §3.1. Figure 5.1 outlines six general risk dimensions that, regardless of the process or technique employed, are important to consider in any risk governance: Context (Domain, Potential Impact), Process (Technical, Business), and Technology (Technique, Data). These are aligned to those previously proposed in technology risk management literature, which provide similar permutations of these six risk categories. Taylor (2012) reviews the relevant literature on technology risks and summarises the key dimensions as: project, relationships, solution, and environment risks [Taylor et al., 2012]. We have used more generic terms (technical process vs. project, business process vs. relationships, technology vs. solution, context vs. environment) to apply to implementations beyond a typical information technology project, to reflect the reality that algorithmic systems are increasingly embedded in a wider array of business functions.

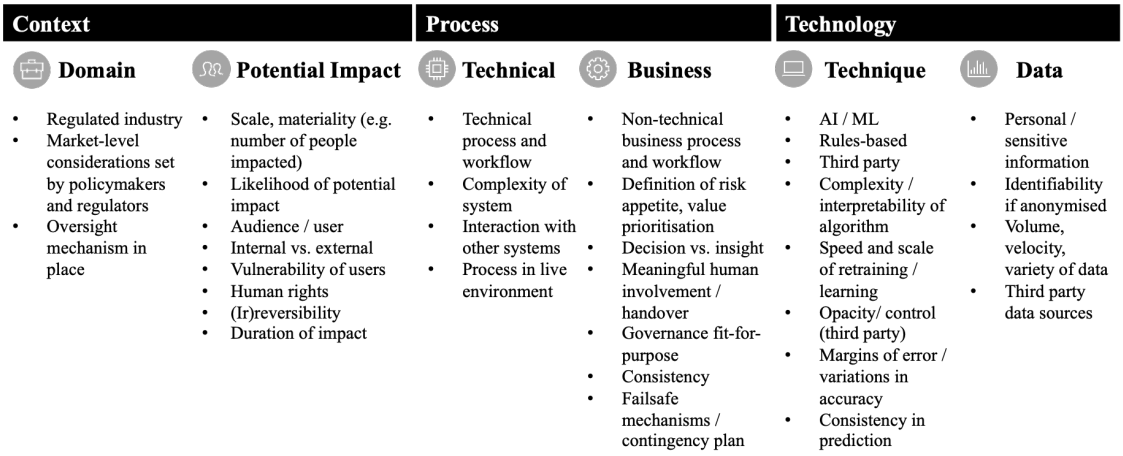


Figure 5.1: Dimensions of risk factors for algorithmic systems

The *process* encompasses both technical and business actions taken in decision-making, e.g. any automated validation checks or stakeholder approvals, while the *technology* is focused on the design, build, and testing of the algorithm (technique) and the data sets used. The *context* includes the domain area, including the relevant regulations beyond data protection, and the potential impact on people, market, and society. Figure 5.1 connects each dimension to the potential risk factors. We do not prescribe these dimensions and factors as 'a' or 'the' comprehensive and complete framework; instead, their role is to demonstrate the nuances and details of potential risks of a system. The risk factors in Figure 5.1 are applicable to all types of algorithms and should be factored in any assessment of system risk. One such factor is the presence of any risks to fundamental human rights. This was discussed in §2.5.1.1.

Based on the risk factors present, there may be different monitoring requirements. For example, a model in a regulated domain area, such as hiring or credit risk, may be subject

to greater scrutiny and oversight. A system that is retrained frequently on large data sets or learns real-time from customer interactions may require automated monitoring and controls to be in place, as manual review is not fit for purpose in keeping up to speed with latest changes. A systematic and standardised approach to monitoring and continuous governance – supplemented by human review and oversight mechanisms – intends to facilitate the detection of potential unknown risks.

5.2 Reviewability: Logging and reporting throughout the lifecycle

Logging and reporting requirements may differ based on the risks and potential impact, as logs should be fit-for-purpose to the risk and proportionate to the level of potential impact. Based on our paper [Cobbe et al., 2021], we discuss the concept of reviewability: technical and organisational record-keeping and logging mechanisms that expose the *contextually appropriate* information needed to assess algorithmic systems, their context, and their outputs for legal compliance, whether they are functioning within expected or desired parameters, or for any other form of assessment relevant to various accountability relationships. While accountability is not a simple challenge with a simple solution, reviewability plays an important role in supporting accountability.

It is important to note that to assess compliance, certain types of reviewers do not need to understand the full inner workings of all technical components of the system. Rather, they would consider the decisions and justifications made by the developer in each of the stages of the development lifecycle, which may include the selection of Key Ethics Indicators, data collection, feature engineering, and testing procedures. For example, an internal audit team should have sufficient understanding of the techniques used to understand their risk factors for the purpose of their review, but it may not be within the scope of the internal audit to independently test and verify the success metrics of the model.

There are several tools in the academic literature that have been introduced to assist in logging and reporting that would provide information necessary for reviewability. First is "datasheets," which details the provenance of data, including its lineage and decisions made throughout its lifecycle—creation, collection, collation,processing, and sharing [Gebru et al., 2021; Singh et al., 2018]. Gebru et al. (2021) propose standardised documentation processes for datasets for the data collector regarding: 1) motivation, 2) composition, 3) collection process, 4) pre-processing/cleaning/labelling, 5) uses, 6) distribution, and 7) maintenance. It is important to consider these aspects of the data sets and their relationship to the system's potential risks. For example, any changes in the incoming data distribution would be logged in the data sheets.

Building on datasheets, 'model cards for model reporting' offer standardised documentation procedures to communicate the performance characteristics of trained ML models [Mitchell et al., 2019]. The model card includes illustrative examples of questions in 9 categories: 1) Model Details, 2) Intended Use, 3) Factors, e.g. demographic or phenotypic groups, environmental conditions, technical attributes, 4) Metrics, 5) Evaluation Data, 6) Training Data, 7) Quantitative Analyses, e.g. of fairness, 8) Ethical Considerations, and 9) Caveats and Recommendations. Model cards include some details on evaluation data and metrics, including performance measures, but models' metrics should explicitly include ethical considerations, such as fairness. We propose explicitly recording Key Ethics Indicators as a part of the model card.

While interventions such as datasheets and model cards are useful, they only provide information about certain stages of an ML system, with a focus on phases of data collection, model selection, and model testing. The non-model processes in the system must be covered as well, including tracking any approvals and sign-offs or documenting advice from experts, such as legal or technology risk teams. Reviewability as a framework encompasses a holistic review of the end-to-end lifecycle of an ML system, including human elements and business processes.

As proposed in our paper [Cobbe et al., 2021], supporting accountability and transparency requires information from systems that is *contextually appropriate*, including what is:

1. relevant to the accountability relationships involved, such as to whom is the account owed and in what format;

2. accurate: correct, complete, and representative;

3. proportionate to the level of transparency required, at the appropriate level of granularity of information and degree of knowledge; and

4. comprehensible by those to whom an account is likely to be owed.

Ensuring reviewability requires fit-for-purpose documentation of relevant information for those with whom the developer has an accountability relationship. As the developer makes decisions that may affect the ethical implications of the model, including any biases duplicated or introduced, these should be clearly documented. This facilitates future review and any required correction or changes to the system design, which is crucial for the monitoring phase. In an *online* learning environment, the retraining cycles should be tracked for any consequential shifts in KEIs. This is even more important for systems with high levels of uncertainty.

5.3 Fairness under uncertainty

Fairness in machine learning has been studied primarily from the perspective of supervised learning, which can be seen in Chapter 4, Figure 4.1, where we have shown that fairness toolkits only handle regression and classification problems. The metrics defined in §2.1 and fairness toolkits introduced in §4.1 are for supervised learning settings. However, many real-life ML applications are online and sequential. Many high-stakes settings require dynamic decisions, rather than static predictions. In domains such as insurance pricing, fraud detection, hiring, and lending, predictions are not evaluated in a one-off mass data set; rather, a decision is made for each individual or a batch of individuals, and the outcome of that decision informs future policies. For example, in insurance, pricing may depend on our updated beliefs about risk. If an area previously deemed to be low flood risk is flooded, the insurance price for that region may increase. In fraud, new patterns may emerge, changing our model. In hiring, organisations may seek to hire similar or different profiles to past hires depending on their performance and depending on availability of a particular profile in the market. Lenders may discontinue mortgage types with high default rates or withdraw mortgage offers based on market conditions, such as pending an increase in interest rates. Therefore, modelling decisions made at a point in time reflects the beliefs and conditions at the time, which may be updated in the future.

In addition, each decision in such a setting is made under uncertainty. The presence of uncertainty, particularly in the feedback loop, is largely overlooked in supervised ML. For example, a bank cannot know whether a denied loan would have been repaid, and it may have less data about previously marginalised and financially excluded populations. An organisation would not know whether an applicant who was rejected would have performed well in the job. An insurer does not know whether a customer would have purchased insurance at a different price point. Such sequential decision-making procedures are more naturally modelled by reinforcement learning (RL) algorithms, which seek to maximise expected rewards over trials by exploring the unobserved space.

In this section, we introduce a taxonomy of uncertainty in sequential decision-making, including "model uncertainty," "feedback uncertainty," and "prediction uncertainty." These uncertainties represent unrealised gains and losses, and we illustrate the potential harms for both the decision-maker and those affected by the decisions of naïve policies that ignore the unobserved space. Outside of this thesis, we are currently working on an algorithm that actively incorporates prediction uncertainty by simultaneously minimising the outcome variance for historically disadvantaged groups and maximising the decision-maker's expected utility.

5.3.1 Legal risks of RL: non-discrimination and equality

Importantly, RL algorithms and subsequent decisions informed by them can pose challenges with regards to non-discrimination laws. The international human rights legal framework, codified in the Universal Declaration of Human Rights and supported by other treaties and documents, establishes the principles of non-discrimination on the basis of certain features such as sex, race, language, or religion [Assembly et al., 1948]. While supervised ML metrics and mitigation techniques have been analysed for their incompatibility with European non-discrimination law [Wachter et al., 2021], limited attention has been given to the legal implications of RL techniques.

RL, however, has the potential, in many ways, to be more problematic than supervised ML from a legal standpoint. There are three contexts in particular in which RL can raise concerns with regards to certain non-discrimination laws: 1) in high-impact decisions, 2) when individual fairness is crucial, and 3) when there are few sequential decisions made in a limited time frame.

RL works through exploration in a stochastic environment, necessarily introducing randomness in decisions. This may involve a nonzero probability of denying a loan to someone who is expected to repay under the model. As [Kilbertus et al., 2020b] note in their article on fair RL, "not all exploring policies may be (equally) acceptable to society". The cost of exploration may be too high, unjustified, unjustifiable, or indeed, illegal in certain domain contexts, such as in criminal justice, employment, and essential financial services. This is the motivation for a proposal for a "semi-logistic" algorithm [Kilbertus et al., 2020b] that combines supervised and reinforcement learning. It represents a guardrail for stochasticity, such that exploration is limited to a specific search space.

Moreover, RL may not achieve individual-level fairness [Dwork et al., 2012], which mandates that "similar" individuals are treated "similarly." Traditionally, US non-discrimination laws and fair lending laws have scrutinised the *inputs* of models, which may be be effective for rules-based systems but less relevant for ML models that rely on complex correlations [Gillis, 2022]. While, again, Gillis (2022) focused on supervised ML, model inputs are even less informative for RL, in which stochastic learning may render individual decisions inconsistent even with the exact same inputs. Indeed, even if RL would lead to a more optimal policy on a group level, *there may be sub-optimal and potentially discriminatory decisions made in the exploration process on an individual level which can have direct legal consequences and liability implications.* Further, in smaller data sets with fewer decision points, due to the time taken for exploration and exploitation in RL, optimal policy may not even be achieved using RL within a time frame that is acceptable to the decision-maker. For example, in Kilbertus (2020) [Kilbertus et al., 2020b], it is not until time step $t = 50$ that stochastic strategies dominate the deterministic policy.

Given the potential legal implications for RL, it may be tempting to assume that RL

is unsuitable for practical applications despite its closer resemblance to practical case studies. However, this is arguably short-sighted given the potential for RL to better handle the uncertainties present in real-life decisions. In fact, our proposal for "fairness under uncertainty" targets exploration specifically to *reduce* unfairness due to uncertainty and provides guardrails on RL for targeted exploration. Our work was motivated by the need for a solution that both accounts for the unobserved decision space and is more aligned to existing legal frameworks and customer expectations then existing work in RL.

5.3.2 Towards fairness under uncertainty for RL

Uncertainty can be more harmful to some groups than others. Previously marginalised groups will often have less data, e.g. due to lack of financial inclusion in lending, and therefore, decision-makers may be less confident about predictions about them. This is a known bias, often termed "representation bias" in §3.2, which can inflate the risk profile of marginalised groups, negatively affecting both profitability and a fair and equitable distribution of financial opportunities.

We propose correcting for this difference in group-level uncertainty through targeted exploration. Intuitively, someone in a minority group *should not be penalised for the model's uncertainty*. We give the minority a "boost" proportionate to the chance that he or she would outperform someone from a majority group. This is meaningfully different from affirmative action, which controversially favours minority groups over majority groups to correct for historical inequalities [Garrison-Wade and Lewis, 2004]. The "boost" in this case is to correct for the false risk inflation for minority due to uncertainties surrounding their prediction.

The exploration-exploitation trade-off we propose in this paper differs from past literature in that it is an active acknowledgement of uncertainty in the ML system, challenging the decision-maker to consider the known unknowns. We discuss this in relation to related work in §5.3.3. The amount of exploration should be proportional to the amount of uncertainty and the decision-maker's risk appetite. Some lenders with a risk appetite close to null may decide to approve only the loans with close to 100% certainty of repayment, thus making defaults extremely rare. This would indicate no appetite for exploration, so the lender may use supervised ML. However, these would be the exception, as risks are inherent in financial decision-making, and most lenders would seek a balance between risk and reward.

For lenders learning from outcomes of past loans to inform future decisions, it should not be naïvely assumed that all denied loans would have defaulted, as this assumption would especially harm marginalised groups. Instead, the assumptions around the current decision boundary should be constantly challenged, taking seriously the differences in uncertainty the decision-maker has about each individual and each sub-group.

Our main contributions are threefold. (1) We propose a taxonomy of uncertainty in fair RL. (2) We formalise the problem of "fairness under uncertainty" for RL, and demonstrate why risk aversion leads to discriminatory policies when decisions produce selective labels. (3) We implement an RL algorithm that outperforms baselines on a range of simulated and real-world datasets.

5.3.3 Related work in uncertainty and RL

There have been few recent works in fair RL. While it represents a step in the right direction in its consideration of stochasticity to counteract feedback uncertainty (the "known unknowns"), we argue that this literature has so far failed to properly acknowledge the nature or consequences of the different layers of *uncertainty* in fair decision-making, described in §5.3.4.

The proliferation of literature on fairness in supervised learning has heavily focused on defining metrics for post hoc model testing or proposing constrained optimisation techniques. These have been discussed with metrics in Chapter 2 and toolkits in Chapter 4. These metrics pointedly only consider the point prediction, rather than the interval around the prediction, and the actual outcome, many of which may be unobserved. When the uncertainty of a predicted outcome is high, when presented with two options with the same average payoff, risk-averse agents (e.g., lenders) may rationally choose to prioritise decisions about which they feel more confident.

Fairness has not received as much attention in RL, although that is beginning to change with the growing awareness that many socially significant applications of machine learning are sequential decision tasks. Early work in this area examined the theoretical properties of algorithms that adhere to strict definitions of individual fairness [Joseph et al., 2016; Jabbari et al., 2017]. More recently, authors have focused on the feedback effects that inevitably emerge in dynamic systems with selective labelling [Lakkaraju et al., 2017; Liu et al., 2018; D'Amour et al., 2020; Kilbertus et al., 2020b], which leads to suboptimal outcomes for all agents. We build especially on the work of Wen et al. (2021) [Wen et al., 2021], who propose methods for fair learning in Markov decision processes (MDPs); however, we expand on their formalisation that fails to account for counterfactual rewards.

Recent contributions have emphasised causal approaches, acknowledging that structural dependencies between variables have important implications for fairness in RL [Zhang and Bareinboim, 2018; Nabi et al., 2019; Creager et al., 2020; Huang et al., 2021]. We argue that this is especially important in the selective label setting, where optimal policies are defined with respect not just to observed data but to all potential outcomes.

Overall, our departure from past literature are in three areas. First, we consider the full range of different types of uncertainties, rather than only tackling feedback uncertainty through RL. Second, we actively account for counterfactual utility: the unobserved gains

and losses. Third, in understanding fairness, we consider the uncertainty interval around each prediction rather than the point prediction alone.

5.3.4 Taxonomy of uncertainty

While scholars have studied subsets of various uncertainties in ML lifecycle, to our knowledge, there is no work to date that holistically examines all types of uncertainties. In this section, we aim to present an uncertainty taxonomy.

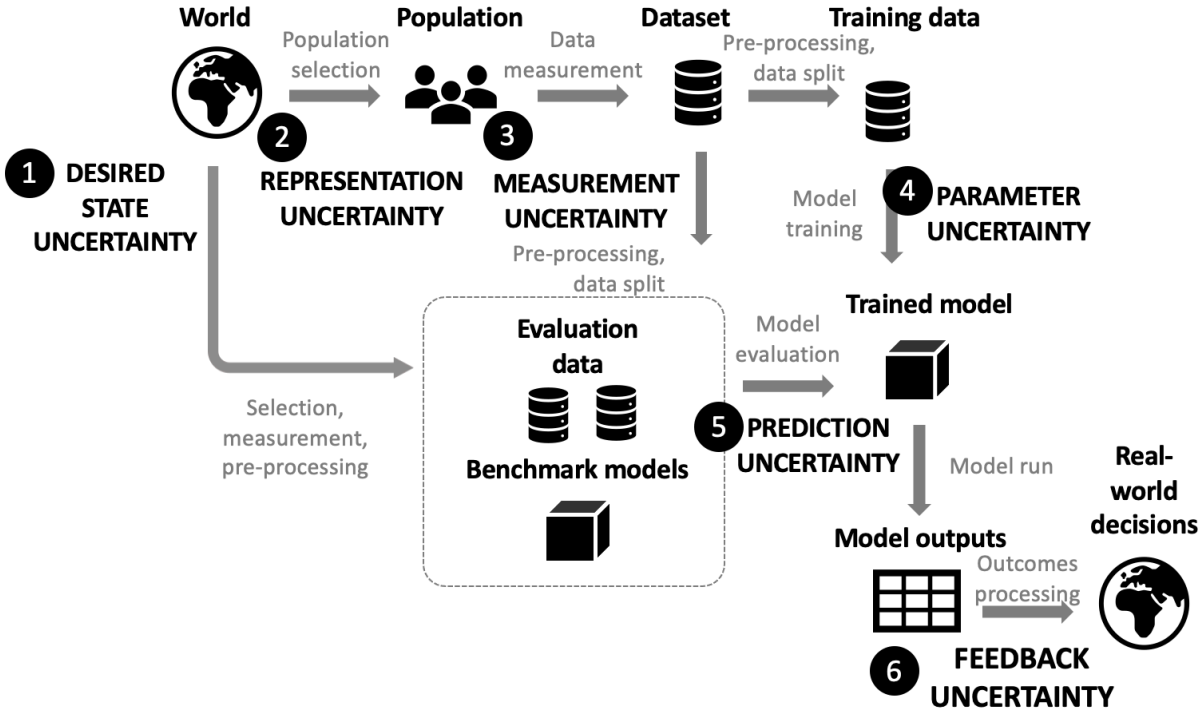


Figure 5.2: Uncertainty taxonomy throughout ML lifecycle

Figure 5.2 shows the five types of uncertainty. Uncertainties 1-4 are "global" uncertainties that affect the model on a systemic level. These should be considered to inform design choices throughout the ML lifecycle. 5 and 6 are "local" uncertainties on an individual or sub-group level, which should be considered alongside the point predictions in the ML algorithm.

While this thesis focuses on fairness in online learning systems best modelled by RL, the uncertainty taxonomy applies broadly to all supervised learning models. This is a contribution on its own, and we have excluded the theory and methods, which are currently in development.

Historically, probabilistic ML literature has not tended to distinguish between inherently different sources of uncertainty, except for more recent references to *aleatoric* and *epistemic* uncertainties [Hüllermeier and Waegeman, 2021; Bhatt et al., 2021]. In the context of ML,

aleatoric uncertainty refers to stochastic variability, such as of a coin toss, and epistemic uncertainty is caused by limitation of the modeller's state of knowledge [Hüllermeier and Waegeman, 2021]. The former gives information on noise or class overlap of the data, and the latter shows where the uncertainty can be reduced by collecting more data in input regions where the training dataset was sparse [Bhatt et al., 2021]. While this framing of reducible (epistemic) and irreducible (aleatoric) uncertainties is useful, there are various sources of both types throughout the ML development process. This section aims to break down each uncertainty into its component sources.

1. World: Desired state uncertainty The fundamental uncertainty on the level of worldview is: how much of the existing inequality is undesirable and should be actively corrected? For example, in lending, discrimination in the job market may increase the actual credit risk of women compared to men. In our past paper [**Lee** et al., 2021] and in Chapter 2, we argue that it is up to the decision-maker to define which types of inequalities are acceptable or unacceptable in any use case, including inequalities in genetics, talent, and socioeconomic ability. To an extent, this is a subjective judgement without a single answer. However, the uncertainty can be reduced through a better understanding of types of inequalities and their impact on the decision-maker's key objectives. Therefore, it can be classified as a type of epistemic uncertainty.

2. Data collection: representation uncertainty To what extent is the data set skewed compared to the target population? "Representation bias" is a known issue among industry practitioners (See §3.2), which is the skewed sampling of the training data set [**Lee** and Singh, 2021a; Holstein et al., 2019]. Representation *uncertainty* refers to the unknown variability of the size and direction(s) of this bias and is a type of reducible (epistemic) uncertainty. The known mitigation technique is to test for representation (e.g. proportion of women in data set) against known population (e.g. proportion of women in the country) and consider additional data collection [**Lee** and Singh, 2021b]. This may be a dynamic uncertainty in the monitoring phase, as the representativeness of a data set may constantly change with each incoming data set. For example, some users may update their privacy preferences to request the deletion of some of their data. If these privacy preferences are associated with the individual's demographic characteristics (e.g. if women are more likely to have greater privacy demands), then this creates new representation bias due to their voluntary exclusion from their data set. However, the demographics of these newly excluded population may not be known, e.g. if the organisation does not collect gender information. Thus, representativeness in these data sets are uncertain as to their reflection of the target population.

3. Feature engineering and selection: measurement uncertainty How well do our variables / features measure what the decision-maker would like to measure? It is well-understood that data sets may contain undesirable proxies of demographic

features [Corbett-Davies and Goel, 2018]. This may include issues of data quality, practical assessment and mitigation techniques for which has been widely studied [Loshin, 2010]. This is a type of reducible (epistemic) uncertainty, as it represents the gap between what is known and what is true.

4. Model training and build – model uncertainty How close are the model parameters to a "true" model? The model built may not be the optimal policy, and a model may exist that better represents the feature relationships. This is a type of reducible (epistemic) uncertainty but differs from measurement uncertainty in its focus on the model type and parameters rather than its training features. Bhatt et al. (2021) [Bhatt et al., 2021] breaks this down further into *model uncertainty* in reference to model parameters and *model specification uncertainty* in reference to the model type. We group them together here because they are a part of the same phase of the ML lifecycle in training and building a model. Dimitrakakis et al. (2019) [Dimitrakakis et al., 2019] addresses this uncertainty through RL. However, their formalisation does not address the remaining uncertainty, and it is unclear how they are compatible with non-discrimination laws, as discussed in §5.3.1. Our proposed approach addresses the model uncertainty through RL that explores the unobserved space in a targeted way that takes into account the decision-maker's risk appetite.

5. Predictions and test data: prediction uncertainty The uncertainty around each prediction (e.g. of probability of repayment) may vary by applicant. There has been work around confidence intervals around each prediction based on the training data, but not in the context of fairness. In particular, uncertainty around the predictions for previously marginalised or excluded groups may be comparatively high, due to representation bias [**Lee** and Singh, 2021b]. This is a key uncertainty we aim to address, by comparing decision boundaries between individuals. At a point in time, this is a type of aleatoric (irreducible) uncertainty because for each applicant, there is a probabilistic density distribution of his or her outcome. However, on a systemic level in an online learning setting, it can also be considered epistemic (reducible) because as the model gains more data about similar applicants, its uncertainty around an applicant could decrease over time.

6. Deployment and retraining – feedback uncertainty Often, the decision-maker's actions determine what data are collected. For example, a rejected job applicant's performance is not measured, and whether a denied loan would have been repaid is unknown. A targeted RL can explore whether the decisions were valid through an analysis of the counterfactual: the expected utility had the decision been different. This uncertainty only exists in online learning settings due to the "known unknowns" in the system. This is slightly more difficult to classify, as it is an epistemic uncertainty in a sense that it is due to the modeller's limited knowledge on the state (or in this case, the counterfactual state) of the world. However, while epistemic uncertainty is characterised by its reducibility by

collecting more data, in this case, the counterfactual state is not knowable for certain. Therefore, this is a type of epistemic uncertainty that is irreducible in practice.

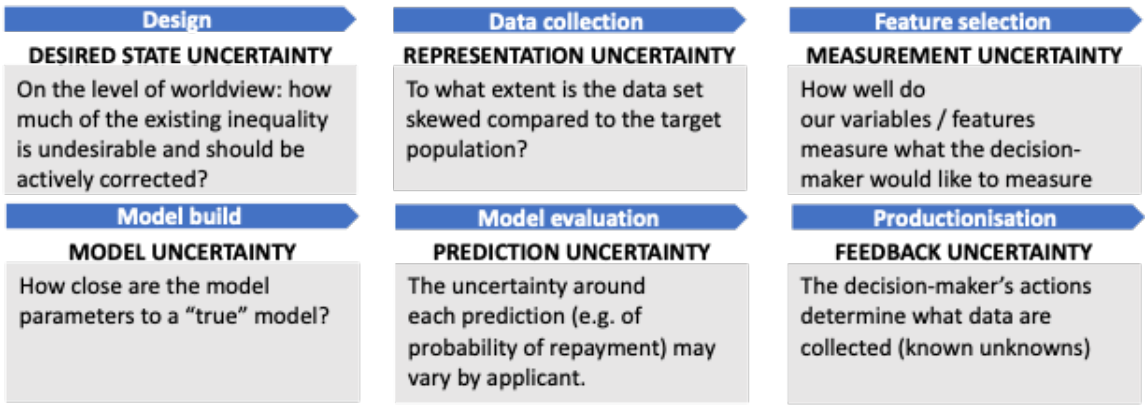


Figure 5.3: Defining the uncertainty taxonomy throughout the ML lifecycle

Each of the global uncertainties should be mitigated at their source: desired state uncertainty by discussion with key stakeholders, representation uncertainty by collecting more data, and measurement uncertainty by fixing data quality issues.

5.3.5 Discussion and future work on fairness under uncertainty

In this section, we have proposed a taxonomy of uncertainty that addresses the end-to-end lifecycle of a supervised ML system. Using visualisations on a simulated data set, we explained the intuition behind the need to address these uncertainties and how this relates to fairness, as under-privileged and marginalised groups tend to have greater local uncertainties in their predictions. We explained why an *online learning* system with high levels of uncertainty is better represented by a hybrid "semi-logistic" model of reinforcement and supervised learning: to enable a more targeted exploration of "known unknown" search spaces in each feedback loop. Our implementation to address these uncertainties currently in progress in collaboration with experts in related fields.

It is important for academics and practitioners in the fair ML community to consider seriously the presence of uncertainty throughout the ML lifecycle and the potential implications. Ignoring the "known unknowns" or the individual-level uncertainties not only potentially puts previously excluded populations at a disadvantage but also results in a worse model performance. Whereas the performance metrics only considering the visible space may improve, the true utility calculation should include unobserved gains and losses as well. Through a holistic understanding of uncertainties in the ML lifecycle, the developer can build a better understanding of the reliability of the model, enabling an informed communication of model predictions to the non-technical stakeholders.

So far, there has been limited work at the intersection of uncertainty in online ML systems and fairness of those systems. Potential next steps in this research include further work on strategies to reduce each of the six layers of uncertainty, which may include non-technical measures such as collecting additional data. Another area is the formalisation and implementation in other domain areas, such as in insurance pricing that should be represented as a regression rather than a classification problem. We hope this work marks the starting point in exploring how we can better incorporate uncertainty considerations in assessing a system's fairness, especially in ensuring previously excluded and marginalised populations are not unfairly penalised for a model's uncertainty about their predictions.

5.4 Key chapter takeaways

This chapter aimed to address the post-go-live monitoring of an ML system to ensure it is still operating fairly and ethically within the pre-set boundaries of KEIs. ML systems differ from static models, such as robotic process automation, in that it learns and adapts from new incoming data. This learning cycle leads to dynamic risks that may change from one time point to another.

To this end, we identified the key risk factors in AI and automated decision-making that would determine the frequency and level of detail required in monitoring. For example, systems in regulated domains, customer-facing systems, and systems with frequent retraining cycles on large data sets would be subject to greater scrutiny, regardless of the type of models being used.

We also outlined the record-keeping and reporting requirements that vary based on the above risk factors. In a *reviewability* framework proposed in our paper [Cobbe et al., 2021], we argue for a fit-for-purpose documentation that is proportionate to the potential impact.

Finally, we proposed a novel uncertainty taxonomy specific to the ML lifecycle and address how these uncertainties may disproportionately affect previously excluded and marginalised groups. In this section, we posit that in some use cases with high levels of uncertainty, it may be sensible to design an RL system that uses the dynamic nature of its learning process to mitigate some of these uncertainties. Monitoring mechanisms and reviewability requirements, however, still apply whether a system uses supervised ML or RL and should still be proportionate to the system's risks.

An end-to-end ML governance does not terminate with deployment. It requires continuous effort with monitoring how the ML system learns over time. Decision-making systems often have "known unknown" spaces, such as whether a denied loan would have defaulted or whether a rejected job applicant would have performed well. Only by closely examining the uncertainties in the system can we challenge our own assumptions in the

model design. With appropriate attention to the risk factors, record-keeping requirements, and uncertainties, the developer can move the ML system toward – not only better predictive performance – but also greater confidence in its fairness and ethics.

Chapter 6

Conclusion

The increasing pervasiveness of ML to inform high-impact decision-making has brought to the forefront the debate around what constitutes a "fair" decision. It is important to remember that many of these concerns over the practicality of "fairness" metrics are not new: in the 1960s and 1970s, the attempts to formalise fairness into statistical metrics fizzled for the same reason: it was not possible to create a broad technical solution to fairness [Hutchinson and Mitchell, 2019]. Despite this, academics have proposed numerous definitions and subsequent formalisations of fairness.

We do not dispute that the ability to quantify fairness, however narrowly, is useful. It allows for benchmarking and comparison between models and can measure the scale of potential impact for any groups. Fairness toolkits recently emerged to make these methods widely accessible, facilitating the implementation of fairness testing methods introduced in academic literature. However, what is crucial to our understanding of an ML system's fairness is the ability to translate these metrics into real-world impact. A solution to a socio-technical problem of fairness cannot be solely technical; technical fixes should be supplemented with changes to people and processes. In complex, real-world ML systems, unfair bias is not a tumor in an otherwise perfect model that can be surgically removed using a "de-biasing" algorithm. The metric must be considered holistically in *context* of all competing objectives of the ML system, including broader ethical considerations.

These fairness metrics, often being both morally irreconcilable and mathematically incompatible with one another, can give the decision-maker conflicting information about whether or not an ML system is fair. In Chapter 2, we link the fairness metrics to the doctrine of ethical philosophy by which they were inspired, pointing out the important gaps between the fairness metric and the original understanding of fairness. What is fair has been a topic of debate throughout human history, hinging on our belief on what types of inequalities are unjust and should be actively remedied. Past work has attempted to bypass the responsibility of this value judgement by arguing it is outside of the scope of a computer scientist's or a developer's responsibilities. *Au contraire*, every decision made by

the ML developer, including the selection of features and performance metrics, has the potential to introduce unintended biases and affect the ethical outcome of the model.

It is true that the decision on what are the model objectives and what trade-offs are acceptable often should lie with the key stakeholders, which may be the board of directors or the business lead, rather than the developer alone. In particular, there may be crucial trade-offs among the model's potential impact and its fairness. In Chapter 2, we also drew on literature in welfare economics to show that in this field, *utility* and potential impact are a critical component of understanding *fairness* from an egalitarian point of view. What matters is not only the distribution of resources but also the size of the resources themselves that are affected by the model.

This is why communication between the developer and the non-technical stakeholders is important, especially in decisions made by the developer that can affect the model. We proposed an approach called "Key Ethics Indicators" to operationalise various ethical and fairness-related objectives into metrics that can be quantified, tracked, and compared between models. These KEIs would be used by the developer in evaluating models and signed off by the relevant stakeholders. In Chapter 3, we also introduced a questionnaire that guides the developer to document potential biases that are introduced throughout the build lifecycle, from data collection methods to feedback mechanisms. In this thesis, we focused on a subset of stakeholders that hold decision-making power around an ML system: primarily on the ML developers, but also the business leaders, the business risk functions, the policy-makers, and the regulators. However, this is not comprehensive of all stakeholders. There is a growing set of important literature that engages with users and communities affected by the ML systems. While this has been out of scope for this thesis, it is important to be cognisant of the incentives, objectives, and preferences of all stakeholders.

Fairness must be understood – not at the algorithmic level – but at the level of the *system*, which includes the process and the people. The questionnaire in Chapter 3 represents our attempt at accounting for the end-to-end pipeline of an ML system in understanding whether there are any unintended biases that result in unfair outcomes. Chapter 4 demonstrated the gaps in existing fairness toolkits, especially in their coverage of the ML lifecycle. We show that coupling the KEI approach and the questionnaire, we are able to uncover the fairness issues and unintended biases in an insurance fraud detection model. Identifying the *source* of each bias also facilitates the design of a targeted mitigation strategy. Bias introduced through language barriers in insurance claim handlers cannot be resolved technically; it is mitigated through hiring multi-lingual staff or using a translator.

Finally, the unique challenge of many ML systems is that they are *online*, learning from previous data points and evolving over time. In particular, online learning algorithms have

"known unknowns" or blind spots. We will not know whether a denied loan would have defaulted or been repaid. In addition, our predictions in ML have a level of uncertainty around them. This uncertainty may be greater in populations that have been previously excluded or marginalised, based on under-representation in the training data set. Intuitively, people should not be penalised due to a model's uncertainty that may derive from past exclusion or past discrimination.

To this end, we proposed a taxonomy of "global" (system-level) uncertainties and "local" (individual-level) uncertainties that apply to all supervised learning models in Chapter 5. We argued that systems with high levels of uncertainty may be best modelled by a hybrid of supervised and reinforcement learning, in which the system "takes a chance" within the uncertainty boundaries through targeted stochasticity. After explaining the intuition using visualised simulated case study, we provided a general formalisation of this method and a pseudocode for implementation.

In the introduction, we started with the reasons why organisations may desire to ensure fairness of their algorithmic decision-making systems. Indeed, this is the underlying assumption in much of the thesis, hence the focus on *unintended* biases rather than malicious agents' purposeful introduction of biases. We also focus on organisation-level decisions and their potential impact, rather than the market-level implications. However, there is scope for further study on whether the existing incentive structure is sufficient to entice organisations to devote valuable time and resources to fairness testing. One 2019 study found that industry practitioners often cite limited time and budget and point out that any fairness investigation happens in their own time, often with limited support for such initiatives from their team or company leadership [Holstein et al., 2019]. Without incentives, there would not be the organisational policies and processes in place to dedicate resources to fairness and ethics more broadly. This is the role of policymakers and regulators: to determine if the industries can self-regulate based on guidelines alone, or whether further laws and regulations are needed to provide additional incentive for action.

In this fast-moving, multi-disciplinary field of ML fairness, we may not have covered all of its vast literature. Our work is not presented as the panacea to fairness issues in ML. Regardless, this thesis purposefully moves away from the false simplicity of technical solutionism and reductionism. By presenting an end-to-end, holistic, and contextualised view of what fairness means throughout an ML lifecycle, we aim to provide a more comprehensive set of methods, approaches, and guidance that integrate with each other to help industry practitioners, researchers, and regulators understand fairness of an ML system in practice.

Works Cited

Bibliography

IEEE standard glossary of software engineering terminology. *IEEE Std 610.12-1990*, 610 (12):1–84, 1990. doi: 10.1109/{IEEE}STD.1990.101064.

Doaa Salman Abdou. Using big data to discriminate charged price in the car insurance industry: Evidence from united states. *Proceedings of Business and Economic Studies*, 2(6), 2019.

Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 252–260, 2020.

Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L Mazurek, and Christian Stransky. Comparing the usability of cryptographic APIs. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 154–171. IEEE, IEEE, 2017.

Jeremias Adams-Prassl, Reuben Binns, and Aislinn Kelly-Lyth. Directly discriminatory algorithms. *The Modern Law Review*, 2022.

Nikita Aggarwal. Law and autonomous systems series: Algorithmic credit scoring and the regulation of consumer credit markets. *University of Oxford Business Law Blog*, 2018.

Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, pages 161–170. PMLR, 2019.

AlgorithmWatch. AI ethics guidelines global inventory, 2019. URL `https://inventory.algorithmwatch.org/about`. Accessed 2022 July 12.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 249–260, New York, NY, USA, 2021. Association for Computing Machinery.

Aristotle and TA Sinclair. *Aristotle: The Politics; Translated with an Introduction by TA Sinclair*. Penguin Books Limited, 1962.

Kenneth J Arrow. Uncertainty and the welfare economics of medical care. In *Uncertainty in economics*, pages 345–375. Elsevier, 1978.

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.

UN General Assembly et al. Universal declaration of human rights. *UN General Assembly*, 302(2):14–25, 1948.

Ari Ball-Burack, **Michelle Seng Ah Lee**, Jennifer Cobbe, and Jatinder Singh. Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 116–128, 2021.

Aaron Bangor, Philip Kortum, and James Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.

D. Baum, J. Scalia, J. Judish, and D. Stute. Supreme Court Affirms FHA Disparate Impact Claims. 2015.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, A Mojsilović, et al. AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

Clement Bellet and Paul Frijters. Big data and wellbeing: An economic perspective. *Ethics of Digital Well-Being: A Multidisciplinary Approach*, pages 175–206, 2020.

Sebastian Benthall and Bruce D Haynes. Racial categories in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 289–298, 2019.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, 2021.

Elettra Bietti. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 210–219, 2020.

Reuben Binns. Human judgement in algorithmic loops; individual justice and automated decision-making. *Individual Justice and Automated Decision-Making (September 11, 2019)*, 2019.

Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 514–524, 2020.

Paula Braveman and Sofia Gruskin. Defining equity in health. *Journal of Epidemiology & Community Health*, 57(4):254–258, 2003.

J Brooke. Usability evaluation in industry, chap. SUS: a "quick and dirty" usability scale, 1996a.

John Brooke. SUS: A quick and dirty usability scale. In *Usability evaluation in industry*. Taylor and Francis, 1996b.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

Cansu Canca. Operationalizing AI ethics principles. *Communications of the ACM*, 63(12): 18–21, 2020.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Aisling Ni Chonaire and Jannna Ter Meer. The perception of fairness of algorithms and proxy information in financial services: A report for the centre for data ethics and innovation. *The Behavioural Insights Team*, 2020.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Jennifer Cobbe, **Michelle Seng Ah Lee**, and Jatinder Singh. Reviewable automated decision-making: A framework for accountable algorithmic systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 598–609, 2021.

Nancy S Cole. Bias in selection. *Journal of educational measurement*, 10(4):237–255, 1973.

Patricia Hill Collins. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. routledge, 2002.

Kate Conger, Richard Fausset, and Serge F Kovaleski. San Francisco bans facial recognition technology. *The New York Times*, 14:1, 2019.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Council of Europe Commissioner for Human Rights. Unboxing artificial intelligence: 10 steps to protect human rights. URL `https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64`.

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In *Proceedings of the 37th International Conference on Machine Learning*, pages 2185–2195, 2020.

Kimberle Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139, 1989.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.

Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. reuters, october 2018, 2018.

Robert Denda, Albert Banchs, and Wolfgang Effelsberg. The fairness challenge in computer networks. In *International Workshop on Quality of Future Internet Services*, pages 208–220. Springer, 2000.

Wesley Hanwen Deng, Manish Nagireddy, **Lee, Michelle Seng Ah**, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. Exploring how machine learning practitioners (try to) use fairness toolkits. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Christos Dimitrakakis, Yang Liu, David C. Parkes, and Goran Radanovic. Bayesian fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):509–516, 2019.

Kevin P Donovan and Emma Park. Perpetual debt in Silicon Savannah. *Boston Review*, 2019.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

Ronald Dworkin. What is equality? part 1: Equality of welfare. *Philosophy and Public Affairs*, 10(3):185–246, 1981.

Margaret L Eaton. *Ethics and the Business of Bioscience*. Stanford University Press, 2004.

Ronel Elul and Piero Gottardi. Bankruptcy: Is it enough to forgive or must we also forget? *American Economic Journal: Microeconomics*, 7(4):294–338, 2015.

Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

European Commission. Communication on artificial intelligence. communication from the commission to the european parliament, the european council, the council, the european economic and social committee and the committee of the regions on artificial intelligence for europe (com/2018/237 final), a. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN`.

European Commission. Guidelines on data protection impact assessment (dpia) and determining whether processing is "likely to result in a high risk" for the purposes of regulation 2016/679, b. URL `http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236`.

European Commission. Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679 (wp251rev.01), c. URL `https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053`.

European Commission Independent High Level Expert Group on Artificial Intelligence. A definition of artificial intelligence: main capabilities and scientific disciplines. report — study of 8 april 2019. URL `https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai`.

European Parliament. European Parliament resolution on automated decision-making processes: ensuring consumer protection and free movement of goods and services (2019/2915(rsp)). URL https://www.europarl.europa.eu/doceo/document/B-9-2020-0094_EN.html.

European Union. EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

Sina Fazelpour and Zachary C Lipton. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 57–63, 2020.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

Avi Feller, Emma Pierson, Sam Corbett-Davies, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, 2016.

Marc Fleurbaey. *Fairness, responsibility, and welfare*. Oxford University Press, 2008.

Marc Fleurbaey. Equality versus priority: how relevant is the distinction? *Economics & Philosophy*, 31(2):203–217, 2015.

Luciano Floridi and Josh Cowls. A unified framework of five principles for AI in society. *Machine Learning and the City: Applications in Architecture and Urban Design*, pages 535–545, 2022.

Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and machines*, 28(4):689–707, 2018.

Ulrik Franke. Rawls's original position and algorithmic fairness. *Philosophy & Technology*, 34(4):1803–1817, 2021.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47, 2022.

Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.

Gemma Galdon Clavell, Mariano Martín Zamorano, Carlos Castillo, Oliver Smith, and Aleksandar Matic. Auditing algorithms: On lessons learned and the risks of data minimization. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 265–271, 2020.

Dorothy F Garrison-Wade and Chance W Lewis. Affirmative action: History and analysis. *Journal of College Admission*, 184:23–26, 2004.

David Gauthier. *Morals by agreement*. Oxford University Press on Demand, 1986.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Talia B Gillis. The input fallacy. *Minnesota Law Review*, 2022.

Barry Goldman and Russell Cropanzano. "Justice" and "fairness" are not the same thing. *Journal of Organizational Behavior*, 36(2):313–318, 2015.

Government of the Netherlands. Strategisch Actieplan voor Artificiële Intelligentie (Strategic action Plan on AI, Policy Brief of 18 October 2019, Government of The Netherlands). URL `https://www.rijksoverheid.nl/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voor-artificiele-intelligentie`.

Ben Green and Lily Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*, 2018.

Jerald Greenberg. A taxonomy of organizational justice theories. *Academy of Management review*, 12(1):9–22, 1987.

David H Guston. The anticipatory governance of emerging technologies. *Applied Science and Convergence Technology*, 19(6):432–441, 2010.

Karen Zita Haigh. AI technologies for tactical edge networks. *Keynote presentation for MobiHoc 2011 Workshop on Tactical Mobile Ad Hoc Networking, May*, 2011.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

Emma Harvey, **Lee, Michelle Seng Ah**, and Jatinder Singh. [**Under review**] practical methods for measuring algorithmic fairness with proxy data. *AI and Ethics*, 2023.

Martie G Haselton, Daniel Nettle, and Damian R Murray. The evolution of cognitive bias. *The handbook of evolutionary psychology*, pages 1–20, 2015.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 181–190, 2019.

Deborah Hellman. Measuring algorithmic fairness. *Virginia Law Review*, 106(4):811–866, 2020.

Emmie Hine and Luciano Floridi. The blueprint for an ai bill of rights: in search of enaction, at risk of inaction. *Minds and Machines*, pages 1–8, 2023.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

W Kuan Hon, Christopher Millard, Jatinder Singh, Ian Walden, and Jon Crowcroft. Policy, legal and regulatory implications of a europe-only cloud. *International Journal of Law and Information Technology*, 24(3):251–278, 2016.

Wen Huang, Lu Zhang, and Xintao Wu. Achieving counterfactual fairness for causal bandit. In *NeurIPS Workshop on Algorithmic Fairness through the Lens of Causality and Robustness*, volume 34, 2021.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58, 2019.

Information Commissioner's Office. Big data, artificial intelligence, machine learning and data protection. *Data Protection Act and General Data Protection Regulation*, 2017. URL `https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf`.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1617–1626, 2017.

Seyyed Ahmad Javadi, Richard Cloete, Jennifer Cobbe, **Michelle Seng Ah Lee**, and Jatinder Singh. Monitoring misuse for accountable 'artificial intelligence as a service'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 300–306, 2020.

Heleen Jenssen, **Michelle Seng Ah Lee**, and Jatinder Singh. Practical fundamental rights impact assessments. *International Journal of Law and Information Technology*, 2022.

Per-Olov Johansson. *An introduction to modern welfare economics*. Cambridge University Press, 1991.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Shelly Kagan. Equality and desert. *What Do We Deserve?: A Reader on Justice and Desert*, page 298, 1999.

Shelly Kagan. *The geometry of desert*. Oxford University Press, 2014.

Faisal Kamiran and Indrė Žliobaitė. Explainable and non-explainable discrimination in classification. In *Discrimination and Privacy in the Information Society*, pages 155–170. Springer, 2013.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

Immanuel Kant and Mary Gregor. The metaphysics of morals. 1996.

Andreas Kaplan and Michael Haenlein. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1):15–25, 2019.

Atoosa Kasirzadeh and Andrew Smart. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 228–236, 2021.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

Frank Kelly. Charging and rate control for elastic traffic. *European transactions on Telecommunications*, 8(1):33–37, 1997.

Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In *Uncertainty in artificial intelligence*, pages 616–626. PMLR, 2020a.

Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. Fair decisions despite imperfect predictions. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 277–287, 2020b.

Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*, pages 4842–4852, 2018.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174, 2018.

James Rufus Koren. What does that web search say about your credit? Jul 2016. URL `https://www.latimes.com/business/la-fi-zestfinance-baidu-20160715-snap-story.html`.

Amit Kumar and Jon Kleinberg. Fairness measures for resource allocation. In *Proceedings 41st annual symposium on foundations of computer science*, pages 75–85. IEEE, 2000.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems 31*, volume 30, 2017.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 275–284, 2017.

Michelle Lee, Luciano Floridi, and Alexander Denev. Innovating with confidence: Embedding AI governance and fairness in a financial services risk management framework. *Berkeley Technology Law Journal*, 34(2), 2020.

Min Kyung Lee and Su Baykal. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*, pages 1035–1048, 2017.

Ben Leo. Mo compare: Motorists fork out £1,000 more to insure their cars if their name is Mohammed. *The Sun*, 2018. URL `https://www.thesun.co.uk/motors/5393978/insurance-race-row-john-mohammed/`.

Kornel Lewicki, **Lee, Michelle Seng Ah**, Jennifer Cobbe, and Jat Singh. [**Under review**] out of context: Investigating the fairness concerns of "artificial intelligence as a service". In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

David Lewis. Causation. *Journal of Philosophy*, 70(17):556–567, 1973.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*, 2018.

David Loshin. *The practitioner's guide to data quality improvement*. Elsevier, 2010.

Tom Lowenthal. Essop v Home Office: Proving indirect discrimination. 2017.

Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–14. Association for Computing Machinery, 2020.

Sandra G Mayson. Bias in, bias out. *Yale Law Journal*, 128:2218, 2018.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL `https://doi.org/10.1145/3457607`.

Gert Meyers and Ine Van Hoyweghen. Enacting actuarial fairness in insurance: From fair discrimination to behaviour-based fairness. *Science as Culture*, 27(4):413–438, 2018.

Microsoft and contributors. Fairlearn, 2019. URL `https://fairlearn.github.io/`.

John Stuart Mill. *On liberty and other essays*. Oxford University Press, USA, 1998.

Christopher J Millard. *Cloud computing law*, volume 2. Oxford University Press Oxford, 2013.

Neha Mishra. Data localization laws in a digital world: Data protection or data protectionism? *The Public Sphere (2016), NUS Centre for International Law Research Paper*, (19/05), 2015.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

Brad A Myers and Jeffrey Stylos. Improving API usability. *Communications of the ACM*, 59(6):62–69, 2016.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, pages 4674–4682, 2019.

Sauradip Nag, Palaiahnakote Shivakumara, Yirui Wu, Umapada Pal, and Tong Lu. New cold feature based handwriting analysis for ethnicity/nationality identification. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 523–527. IEEE, 2018.

Arvind Narayanan. Tutorial: 21 definitions of fairness and their politics, 2018. URL `https://www.youtube.com/watch?v=jIXIuYdnyyk`.

Anthony Nardone, Joey Chiang, and Jason Corburn. Historic redlining and urban health today in us cities. *Environmental Justice*, 13(4):109–119, 2020.

OECD.AI Policy Observatory. National strategies, agendas and plans. URL `https://oecd.ai/dashboards/policy-instruments/National_strategies_agendas_and_plans`.

United States. Executive Office of the President and John Podesta. *Big data: Seizing opportunities, preserving values*. White House, Executive Office of the President, 2014.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

Derek Parfit. *Equality or priority*. University of Kansas, Department of Philosophy, 1991.

Charlie Parker, Sam Scott, and Alistair Geddes. Snowball sampling. *SAGE research methods foundations*, 2019.

Edmund C Penning-Rowsell. Flood insurance in the UK: a critical perspective. *Wiley Interdisciplinary Reviews: Water*, 2(6):601–608, 2015.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

Robin A Prager et al. *Determinants of the locations of payday lenders, pawnshops and check-cashing outlets*. Federal Reserve Board Washington, DC, 2009.

Anya ER Prince and Daniel Schwarcz. Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev.*, 105:1257, 2019.

Ariel Procaccia. AI Researchers are pushing bias out of algorithms, 2021. URL `https://www.bloomberg.com/opinion/articles/2019-03-07/ai-researchers-are-pushing-bias-out-of-algorithms`. Accessed: 22 September 2022.

John Rawls. *A Theory of Justice*. Harvard University Press, 1999.

John Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001.

Stuart Russell and Peter Norvig. *Artificial intelligence: a modern approach*. 2002.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018a.

Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018b.

Samuel Scheffler. *The rejection of consequentialism: A philosophical investigation of the considerations underlying rival moral conceptions*. Oxford University Press, 1994.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.

Amartya Kumar Sen. *Inequality reexamined*. Oxford University Press, 1992.

Jatinder Singh, Jennifer Cobbe, and Chris Norval. Decision provenance: Harnessing data flow for accountable systems. *IEEE Access*, 7:6562–6574, 2018.

Rebecca Kelly Slaughter, Janice Kopec, and Mohamad Batal. Algorithms and economic justice: A taxonomy of harms and a path forward for the federal trade commission. *Yale JL & Tech.*, 23:1, 2020.

Kacper Sokol, Raul Santos-Rodriguez, and Peter Flach. FAT Forensics: A python toolbox for algorithmic fairness, accountability and transparency. *arXiv preprint arXiv:1909.05167*, 2019.

Christina Starmans, Mark Sheskin, and Paul Bloom. Why people prefer unequal societies. *Nature Human Behaviour*, 1(4):1–7, 2017.

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9, 2021.

Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0), 2023-01-26 05:01:00 2023. URL `https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225`.

Hazel Taylor, Edward Artman, and Jill Palzkill Woelfer. Information technology project risk management: bridging the gap between research and practice. *Journal of Information Technology*, 27(1):17–34, 2012.

Michelle Seng Ah Lee and Luciano Floridi. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds and Machines*, 2020.

Michelle Seng Ah Lee and Jat Singh. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021a.

Michelle Seng Ah Lee and Jatinder Singh. Risk identification questionnaire for unintended bias in machine learning development lifecycle. *Proceedings of the AI, Ethics, and Society Conference*, 2021b.

Michelle Seng Ah Lee and Jatinder Singh. Spelling errors and non-standard language in peer-to-peer loan applications and the borrower's probability of default. In *In Proceedings of Credit Scoring and Credit Control Conference XVII*, 2021c.

Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI and Ethics*, pages 1–16, 2021.

Michelle Seng Ah Lee, Jennifer Cobbe, Heleen Janssen, and Jatinder Singh. Defining the scope of AI ADM system risk assessment. In *Research Handbook on EU Data Protection Law*. Edward Elgar Publishing, 2022.

Michelle Seng Ah Lee, David Watson, , and Zhe Feng. [IN PROGRESS, PENDING SUBMISSION] Fairness under uncertainty in sequential decisions. In *Proceeding of the 39th IEEE International Conference on Data Engineering (ICDE)*, 2023.

Joe Tomlinson. Quick and uneasy justice: An administrative justice analysis of the EU settlement scheme. URL `https://publiclawproject.org.uk/wp-content/uploads/2019/07/Joe-Tomlinson-Quick-and-Uneasy-Justice-Full-Report-2019.pdf`.

G Tripepi, KJ Jager, FW Dekker, C Wanner, and C Zoccali. Bias in clinical research. *Kidney international*, 73(2):148–153, 2008.

Michael Carl Tschantz. What is proxy discrimination? In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea*, 2022.

Andrew Van Dam. Searching for images of CEOs or managers? the results almost always show men. *The Washington Post*, 01 2019.

Michael Veale and Frederik Zuiderveen Borgesius. Demystifying the draft EU Artificial Intelligence Act—analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4):97–112, 2021.

Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

Neil Vigdor. Apple card investigated after gender discrimination complaints. *The New York Times*, 2019.

James Vincent. *Beyond measure: the hidden history of measurement.* Faber & Faber, 2022.

Matthijs Vincent and ManyOthers. scikit-fairness, 2019. URL `https://github.com/koaning/scikit-fairness`.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *W. Va. L. Rev.*, 123:735, 2020.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, 41:105567, 2021.

Stephen Walli, Dave Gynn, and Bruno von Rotz. The growth of open source software in organizations. *Publication Report. Optaros Inc*, 2005.

Michael Walzer. Spheres of justice. *A Defense of Pluralism and Equality, New York, Basic*, 1983.

Min Wen, Osbert Bastani, and Ufuk Topcu. Algorithms for fairness in sequential decision making. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1144–1152, 2021.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

Lauren E Willis, Olatunde C Johnson, Mark Niles, and Rigel Christine Oliveri. Comments to HUD re: Fr-6111-p-02, HUD's implementation of the Fair Housing Act's disparate impact standard. *Loyola Law School, Los Angeles Legal Studies Research Paper*, (2020-34), 2019.

Laurie F Wurster, Bob Igou, and Zeynep Babat. Survey analysis: overview of preferences and practices in the adoption and usage of open-source software. *Gartner Group (G00210068)*, pages 1–28, 2011.

Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*, pages 563–574. Springer, 2019.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making — the causal explanation formula. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Yongfeng Zhang, Xu Chen, et al. Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval*, 14(1):1–101, 2020.

Minhaz Zibran. What makes APIs difficult to use. *International Journal of Computer Science and Network Security (IJCSNS)*, 8(4):255–261, 2008.

Claire Zillman. Fortune 500 female CEOs reaches all-time record of 33. *Fortune*, 05 2019.

Appendix A

Bias in Model Development Lifecycle Questionnaire

This questionnaire will go through an assessment to identify the potential unfair and unintended biases and discrimination in a model development lifecycle.

It is intended for initial risk identification to facilitate subsequent evaluation, quantification (likelihood/impact), and mitigation. By identifying the potential source of the bias risk, the questionnaire will allow for a more targeted design of a mitigation strategy.

The questionnaire is not intended to be a comprehensive, definitive standard for bias risk assessment; rather, it provides a starting point to further adapt and extend it to be customised to the use case and domain area. Further guidance documents would be developed for practical implementation.

This risk evaluation stage may be used to assess the trade-offs in the model and justify its usage to key stakeholders, both internal (e.g. board) and external (e.g. customers, regulators). It may be used internally by the model development team with input from others, e.g. legal risk teams, by the internal audit / model validation team, or externally for an independent third-party assessment of the ethical risks of the model.

A. Background information

- **A.1 Model description**: Describe a model you would like to assess for potential unfair and discriminatory bias. Ex) a supervised machine learning model to predict whether a mortgage loan will default

- **A.2 Positive impact**: overall**: What positive impact can this model have on the target population? This may include an explicit social impact as intended by the model (e.g. building a computer vision model to read sign language to increase service accessibility), or it could be associated with the efficiency of the business

and market that trickles down into a positive result for the consumers, e.g. more precisely predicting default risk can help prevent unaffordable loans being approved.

- **A.3 Positive impact: performance** What is the benefit of higher accuracy / precision for the target population? Ex) better credit risk evaluation model leads to greater financial inclusion, better hiring algorithm leads to overall higher employee performance / reduction in attrition

- **A.4 Positive impact: operationalisation** Can these objectives be measured and quantified? If yes, list how they can be formalised. Ex) unaffordable loan approval can be measured based on false negative rate (i.e. loans predicted to be repaid but defaulted), and greater financial inclusion can be measured as the total amount of loans given out

- **A.5 Negative impact: allocative harm** What are any potential allocative harms (withholding of opportunities / resources)? Ex) model may be more likely to give loans to certain groups, e.g. race and gender, which would replicate and widen the societal inequalities

- **A.6 Negative impact: representational harm**: Is there any representational harm (diminished identity)? Ex) for an image search algorithm for "CEO", returning more men than women reinforces the bias in identity

- **A.7 Negative impact: representational harm operationalisation** Is there any representational harm (diminished identity)? Ex) for an image search algorithm for "CEO", returning more men than women reinforces the bias in identity

- **A.8 Negative impact: fundamental rights**: Are there any fundamental rights at stake? Ex) right to self-determination, liberty, due process of law, freedom of movement, privacy, freedom of thought, freedom of religion, freedom of expression, right of peaceful assembly, right to freedom of association

- **A.9 Negative impact: operationalisation**: Can these objectives be measured and quantified? If yes, list how they can be formalised. Ex) unaffordable loan approval can be measured based on false negative rate (i.e. loans predicted to be repaid but defaulted), and greater financial inclusion can be measured as the total amount of loans given out

- **A.10 Relevant regulations and laws (e.g. discrimination)**: What are the relevant regulations and laws that can help frame the risk assessment? For example, there are various anti-discrimination legislations and court cases. The definition of unlawful discrimination and the definition of a legally protected characteristic (e.g.

race, gender) may vary across jurisdictions, e.g. as defined in the 2010 Equality Act in the UK, and by domain area, e.g. Equal Credit Opportunity Act in the US. Seek out relevant guidance from official documents and from internal legal / regulatory risk teams.

Based on the answers above, think critically about the model's practical and ethical objectives, and when they may be in conflict with one another, understanding the prioritisation of and trade-offs between these objectives.

Below questions will align to each stage of the model development lifecycle. Answering yes indicates a bias risk that must be addressed, mitigated, and/or justified.

B. Design: historical/external bias

- **B.1 History of discrimination**: Is there documented historical discrimination in the domain area against a protected class, as defined in A.10? Ex) academic studies demonstrate lower mortgage approval rates for racial minorities in the US, especially black and Hispanic applicants. These are sub-groups to which special attention must be paid in testing for impact of the model.
  - age
  - disability
  - gender reassignment
  - marriage or civil partnership (in employment only)
  - pregnancy and maternity
  - race
  - religion or belief
  - sex
  - sexual orientation
  - other
- **B.2 Acceptable vs. unacceptable inequalities**: For that group, select which of the following layers of inequality is a justifiable source of differences in outcome. For example, race and gender may be relevant to differential medical diagnoses, and talent / education may be relevant to recruiting and hiring algorithms. Is there any disagreement among relevant stakeholders?
  - Disability

  - Race
  - Age
  - National origin
  - Socioeconomic status
  - Talent/ education
  - Personality traits
  - Preferences
  - Culture
  - Discrimination in related markets (e.g. for a credit risk model, any inequality resulting from discrimination in the job market)
  - Other

- **B.3 Potential proxies**: Identify the features in your data may be associated with the unjustifiable sources of differences in outcome, e.g. postcode with race or income with gender. Detailed tests would be undertaken in the feature engineering stage. Is there sufficient rationale for including these features, e.g. well-founded causal relationship to an outcome of interest (e.g. income to risk of default) or feature that is within the individual's control and transparently disclosed (e.g. history of paying bills on time)?

- **B.4 World as-is vs. ideal**: Is there any misalignment between the ground truth (world as-is) and the organisation's values? For example, there may be more male senior executives, but the organisation's objective is to have stronger female representation in leadership.

C. Data collection: Representation bias

- **C.1 Selection bias**: Is the marketing / targeting / data collection strategy returning a non-representative sample of the population? Ex) is the mortgage company advertised in majority-white neighbourhoods, or is the recruiting firm only active at top universities?

- **C.2 Subjective recorded features**: Are any of the recorded features affected by human judgment? Ex) the data set may include the interviewer's scores on the candidates' performance

- **C.3 Third party**: Are any of the recorded features produced by a third party data set or model? Ex) the credit scores may be provided by a specialist agency, or an open source data set on university rankings may be used in a hiring model

- **C.4 Known unknown**: Is any ground truth of actual outcomes unknown? Ex) whether denied loans would have defaulted is unknown
- **C.5 Sample size**: Is there insufficient sample in any subgroup of interest (especially those in B.1) for this analysis? Ex) only 1% of applicants are Native Americans

D. Feature engineering: measurement bias

- **D.1 Different measurements**: Are there differences in the measurement process between groups for either input features or the target outcome? Ex) high-minority neighborhoods are more frequently patrolled, leading to higher arrest rates
- **D.2 Different data quality**: Are there differences in the data quality between groups? Ex) schools in poor districts have lower quality recorded data on student performance
- **D.3 Subjective engineered features**: Are there any features added by the model developer that could it be affected by his/her judgment? Ex) the data scientist added flags of what he/she considers an important feature from a job application, e.g. "participated in university extracurricular activities" or "held a leadership position"
- **D.4 Test for proxies**: Are there proxies of outcome that may be also proxies of a protected group membership, especially those with a history of discrimination in this domain area? Refer back to answers from B.3 on whether the inclusion of these proxies is justifiable. If the protected features are available, correlation tests are recommended. Ex) Job type and car value are associated with both auto insurance risk and gender. Car value may be justifiably included due to its causal relation to risk of accident-related claims, but job type may not be fully justifiable, e.g. if a male-dominated job titles, e.g. barbers, are charged different prices to female-dominated job title, e.g. hair stylists. The modeller may consider mitigating the gender bias by using a newly engineered feature more closely associated to the outcome of interest, e.g. typical amount of car usage in job, rather than directly using the job title.
- **D.5 Measurement accuracy**: Is there any mis-match between the measurement and what the model intends to track? Ex) arrest is not equivalent to crime rate, final grades are not equivalent to student success

E. Model build and training: aggregation bias

- **E.1 Heterogeneous groups**: Are the populations heterogeneous in a way that a single model cannot account for all subgroups? (See: Simpson's paradox) Ex) Medical diagnosis algorithm should be different for men and women given their different body compositions

- **E.2 Heterogeneous mechanisms**: Are there other heterogeneous mechanisms in play that are being inaccurately aggregated that may be associated with protected features? Ex) differences in behavior across products, different time periods, different data sets, etc.

F. Model evaluation: evaluation bias

- **F.1 Trade-offs**: Identify all trade-offs on objectives identified for all available models. All objectives should be quantified into metrics where possible to enable model comparison. Is there any tension between key objectives? Ex) mapped the trade-off between financial inclusion and minority race denial rates for mortgage lending for 10 versions of predictive models, but there is no obvious winner

- **F.2 Objective coverage**: Are there any gaps in the metrics' coverage of all measurable objectives related to positive and negative impacts on the target population? Ex) The assessment of mortgage default prediction algorithm covers unaffordable loans (false positive), financial inclusion, minority race denial rate, but explainability should be qualitatively assessed

- **F.3 Metric alignment with values**: Is there any mis-alignment of your model performances with the relative importance of False Positives vs. False Negatives? Ex) those predicted to repay but defaulted represent unaffordable loans / cost to the company, and those predicted to default but would have repaid represent missed opportunity / allocative harm

- **F.4 Metric over-fitting**: Is there a metric the model may be over-fitting to? Ex) the main credit risk evaluation accuracy metric

- **F.5 Sub-group performance disparity**: Is there any difference in model performance, measured on on all protected subgroups, especially those identified in B.1? Be sure to test all subgroup combinations, e.g. intersectional discrimination in race-gender. Ex) the model has similar error rates for men and women and for black and white applicants, but it has high error rates for black women

- **F.6 Confidence**: Is there any barrier to the confidence intervals being accepted and understood by the relevant stakeholders? Ex) Especially if a sub-group population is under-represented, they may have a larger confidence interval around their predictions

G. Model productionisation and monitoring: deployment bias

- **G.1 System**: Is the model a part of a complex sociotechnical system, e.g. interconnected models or embedded in human processes? Ex) A CV-scoring algorithm may feed into a candidate's evaluation system, which should also be assessed holistically beyond the ML model

- **G.2 Feedback**: Are there any gaps in the human feedback mechanism for any errors? Ex) A human reviewer reads a sample of machine transcriptions to identify any errors and retrains the algorithm with the corrections, but high-cost errors may be missed

- **G.3 Robustness to external changes***: Test the model for robustness to any external changes, e.g. shifts in policy, dramatic changes in input data, etc. Are there any concerns on the organisation's ability to monitor its performance in case of any external change? Ex) There is a monitoring mechanism in place to alert the team if there is a significant change in the distribution in the input data beyond a pre-defined threshold

- **G.4 Bias reinforcement**: Can the feedback loop be reinforcing any existing biases? Ex) if loans predicted to default are denied. Is there any user interaction with the output? Ex) user clicking on links recommended by the algorithm

This questionnaire can also be found at `https://github.com/michelleslee/bias_in_lifecycle/`.