Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle

Michelle Seng Ah Lee, Jatinder Singh
Compliant & Accountable Systems Group
Department of Computer Science & Technology
University of Cambridge, UK
(michelle.sengah.lee,jatinder.singh)@cl.cam.ac.uk

ABSTRACT

Unintended biases in machine learning (ML) models have the potential to introduce undue discrimination and exacerbate social inequalities. The research community has proposed various technical and qualitative methods intended to assist practitioners in assessing these biases. While frameworks for identifying the risks of harm due to unintended biases have been proposed, they have not yet been operationalised into practical tools to assist industry practitioners.

In this paper, we link prior work on bias assessment methods to phases of a standard organisational risk management process (RMP), noting a gap in measures for helping practitioners identify bias-related risks. Targeting this gap, we introduce a bias identification methodology and questionnaire, 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.

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

CCS CONCEPTS

• **Software and its engineering** → **Risk management**; *Software development process management*; • **General and reference** → *Metrics*; • **Social and professional topics** → *User characteristics*; *Computing / technology policy*; Testing, certification and licensing; • **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

risk identification, questionnaire, risk management, algorithmic bias, algorithmic fairness, fair ML

AIES '21, May 19–21, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8473-5/21/05.
https://doi.org/10.1145/3461702.3462572

ACM Reference Format:
Michelle Seng Ah Lee, Jatinder Singh. 2021. Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462572

1 INTRODUCTION

As machine learning (ML) models are increasingly used to inform the making of high-impact decisions, there is greater scrutiny on the potential for their predictions to reflect and exacerbate patterns of societal inequalities, unfair discrimination, and exclusion. This has led to a surge in related academic work, from technical methods to define and quantify fairness to open-source implementations of fairness tests to qualitative checklists and logging templates. However, it is currently unclear how these tools and methods best fit into an end-to-end enterprise risk management framework for their practical usage in industry.

1.1 Related work

Studies have consistently shown that practitioners struggle to integrate proposed tools and methods into their existing processes [19, 27]; however, there have been efforts in recent work to move from the conceptualisation of fairness/bias considerations, e.g. definitions, trade-offs, and frameworks, to their operationalisation into tools and methods. These include technical (fairness tests, fairness toolkits) and qualitative methods (checklists, logging templates). Mathematical fairness tests (e.g. [12, 18, 23]) formalise definitions fairness into a metric to be calculated for each model. Fairness toolkits [2, 31, 38, 48, 50] implement these tests, often with some form of user interface and open source code to facilitate their use. Checklists, inspired by those in other domains (e.g. aviation and medicine), have been designed to give practitioners concrete action points [11, 29]. Logging templates, such as Model Cards [32] and Datasheets [15], have been proposed to record important information about data and model being used. All of these methods aim to operationalise past work in fairness and bias frameworks into tools that may be used in real-world contexts.

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 (2019), 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. In each stage of model development, there are decisions made that could result in skewing of the

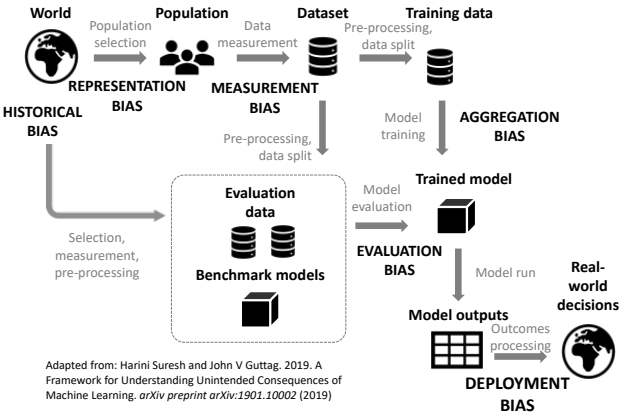

Figure 1: Bias in ML development lifecycle

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 define six categories of biases that may cause unintended harm:

(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 [30, 35]. The ML development lifecycle involves a series of decisions from evaluation methodology to model selection that can lead to unwanted effects (illustrated in Figure 1). As such, instead of "fairness," we refer to *unintended bias* with an eye to any aspects of the data, model, and processes in the lifecycle that may result in negative impact, especially on previously marginalised groups.

1.2 Motivation

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 [45]. 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 risks in their particular context and domain area [19]. Such difficulties for practitioners remain despite the emergence of fairness toolkits, in part due to the tools' limited coverage of ML lifecycle and the confusion on how such methods integrate with organisational processes [27].

General risk management frameworks exist to guide organisations across industry on how to govern and mitigate risks as a part of day-to-day business processes, which are useful in understanding best practices in a typical risk management process. However, scholars have little referred to these well-established frameworks in proposing methods to address the risk of biased models. Given the consensus among empirical studies in risk management on the importance of *proactivity* [42, 51] and *formalisation/standardisation* [49, 51], it is important that bias risks methods are *integrated* into broader organisational risk processes. For example, should fairness toolkits, checklists, and logs all be used in an ML lifecycle; if so, at what stage? Understanding how these methods fit into business processes is essential for their widespread adoption.

To this end, we map the related work to relevant *risk management process* (RMP) phases to demonstrate how they may be used in industry. This is not only a contribution on its own, but also helps place the scope of our paper: we identify a relative gap in the risks identification of potential unintended and harmful biases in ML, for which frameworks have been proposed but the methods never concretely or fully operationalised. Our second contribution aims to help close this gap by proposing a new risk assessment questionnaire, a practical method to be used to identify the *sources* of downstream harm – the biases that result in unintended, potentially unfair outcomes. This risk identification process is built on recent work describing frameworks for unintended biases [30, 35, 41]. We illustrate the questionnaire's usage in practice through a real-life case study: biases in an insurance fraud prediction algorithm. Finally, we verify the potential of the questionnaire through a survey of industry practitioners, highlighting their practical requirements and preferences relevant to development of any future methods.

2 BIAS METHODS AND ENTERPRISE RISK MANAGEMENT

ML practitioners claim to require flexible tools that integrate well with organisational processes [19, 27, 45]. In this section, we map existing methods to a standard risk management framework in order to facilitate the integration of these methods into a typical enterprise governance process. This mapping contributes to present literature by bringing together the international risk standards and the methods proposed for bias risk management, and placing the scope of this paper in relation to previously proposed methods and tools.

Organisational risk assessment is an iterative process [9]. As a standardised practice in cross-industry business contexts, it typically operates to identify key concerns and risk mitigation strategies to maintain residual risk at levels acceptable to the organisational risk appetite. This applies to all activities beyond technological systems. Technology risk frameworks and international standards include: ISO/IEC 27001 standard for information security, IEEE 1540-2001 for software lifecycle processes, and COSO framework [8].

This paper focuses on RMP outlined in ISO 31000, a family of risk management standards codified by the International Organization for Standardization (ISO). It is more generic than the ISO/IEC 27001 or IEEE 1540-2001, encompassing risks beyond information security and software engineering. This is more appropriate because bias

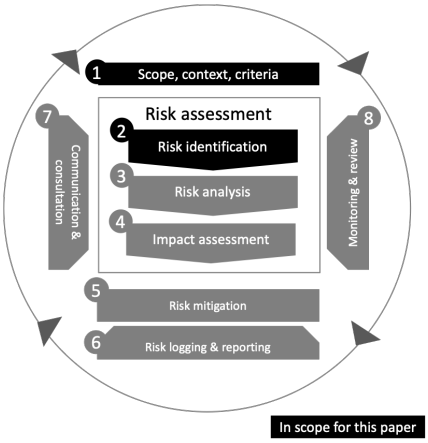

Figure 2: ISO31000:2018 Risk Management Process (RMP)

risks apply across business functions beyond the technical team, e.g. legal and reputational risk teams. It is internationally recognised as the "gold standard" and was widely adopted around the world by most G20 countries, representing the "collective wisdom... on what good risk management looks like" [9]. We focus on ISO 31000 over any software/security-specific frameworks to account for the breadth of potential risks that go beyond the technical challenges, especially societal impact and regulatory considerations. ISO/IEC CD 23894 (Artificial Intelligence Risk Management) and ISO/IEC DTR 24027 (Bias in AI systems and AI aided decision making) are under development, aiming to provide a more targeted risk management guidance for AI/ML technologies. However, ISO 31000 forms the foundational basis for all of its more specific guidance, including new standards [9], and by aligning the literature to the more generic framework, we ensure our work is widely applicable to organisations that are already using ISO 31000.

2.1 Mapping bias risk methods to ISO 31000 processes

As ISO 31000 is a globally recognised benchmark for all types of organisations and practices [9], its guidelines shed light on how organisations typically design RMPs. As such, we map the related work that proposes bias-related operational and practical tools to each of the RMP phases in ISO 31000 (Fig. 2). We briefly discuss some of their limitations in fulfilling the requirements of an *end-to-end* (from assessment to mitigation to monitoring throughout a model lifecycle) RMP – especially those that highlight the importance of a targeted bias risk identification methodology. The objective is not to compile a comprehensive catalogue of relevant literature, but rather to provide indicative examples of the diverse types of methods proposed to tackle bias risk and aligning them to a defined RMP stage.

2.1.1 (1) Scope, context, criteria. Case studies have demonstrated and analysed the risks of unfair bias in areas such as criminal risk prediction [17], health care provisions [37], and mortgage lending [25]. However, more work is needed to contextualise fairness and bias considerations in domains where it is not simple to define and quantify metrics; practitioners cited chatbots and web search as areas with limited guidance [19]. Studies on practitioner needs showed they struggle to adapt the methods from academic literature and in software toolkits to their specific use cases. One states "the predominant mode of development [in research] often involves characterising a problem in a way that might often be at odds with the real world context" [45], a sentiment echoed in subsequent work engaging practitioners [19, 27].

2.1.2 (2) Risk identification. Scholars have proposed mathematical methods (e.g. Dwork et al., Hardt et al., Kusner et al.) to formalise and test for a particular definition of "fairness" in ML. These definitions can be incompatible with one another [22, 36], prompting work distinguishing between them [34, 47]. These techniques assume that fairness can be mathematically operationalised, a view often criticised as overlooking the societal and historical contexts [16, 39].

While these mathematical fairness tests may identify *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." There have been critiques that these definitions give little information or guarantee on model fairness and that there should be a more systematic method for identifying and mitigating the risk of unfairness [7] – a difficult task where there are competing definitions. In particular, 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" [19]. These would not be identified in the mathematical tests and require a qualitative identification.

Frameworks have been introduced categorising these types of unintended biases to *make explicit these potential "blind spots"* [30, 35, 41]. They formally define the types of biases that may affect the outcome in undesirable ways, addressing the full, end-to-end ML development lifecycle.

2.1.3 (3) Risk analysis. Some have implemented the fairness tests mentioned above into fairness toolkits and added visualisations of test results and user interface to enable model and data interrogation [2, 31, 38, 48, 50]. These represent advances in making the fairness formalisations from academia accessible to practitioners. Despite this, a recent study of open source fairness toolkits reports that practitioners find the toolkits (i) difficult to understand, (ii) challenging to adapt to their use case and integrate into their processes, (iii) limited in their coverage of the development pipeline, and (iv) unclear on the potential mitigation strategies [27]. A more effective risk identification process throughout the ML pipeline could facilitate a more targeted analysis to quantify and understand the risk and its source.

2.1.4 (4) Impact assessment. In assessing the real-life implications of a model, several approaches have been proposed, from trade-off analyses to long-term simulations to qualitative assessments. Unintended bias is often framed as a trade-off between (i) accuracy and the benefits associated with a higher-performing algorithm and (ii) "fairness" and the potential discrimination or exacerbation of inequalities and societal prejudice. Such trade-off analyses have

been formalised in criminal justice and in credit risk [22, 25]. Some scholars have found the impact assessment may contradict the fairness tests mentioned in stages 2 and 3 (risk identification and analysis); a study of the long-term impact of a "fair" ML model was shown to harm the historically disadvantaged sub-group it intended to protect [28]. In addition to these quantitative impact assessments, several qualitative impact assessments have been introduced that are specific to algorithmic bias, especially Human Rights Impact Assessment and Ethical Impact Assessment [20].

2.1.5 (5) Risk mitigation. Some fairness toolkits offer "de-biasing" techniques [2, 50], including pre-processing (removing unfairness from data) (e.g. [13]), in-processing (adding constraints during training) (e.g. [3]), and post-processing (correcting unfairness in the predictions) (e.g. [21]). These techniques have been critiqued for over-simplifying the socio-technical contexts of a bias and for disregarding the potential real-life impact [7, 28].

While these fairness tests and technical de-biasing may be useful in certain settings, they should be supplemented by practical and contextual guidance for any corresponding non-technical mitigations. This involves breaking down the technical and organisational elements of ML-driven decision-making [6]. For example, many mitigations involve people and processes rather than the model itself, e.g. training human data labellers or collection of more diverse data sets [25]. Practitioners have found these "de-biasing" methods are insufficient in addressing risks and often incompatible to their own domains, precisely because the methods only frame bias in a narrow, technical sense, ignoring the process and context around the technical implementation [19]. More effective risk identification processes would help diagnose the sub-populations and types of biases to inform the appropriate mitigation strategy, whether technical or people/process-driven.

2.1.6 (6) Risk logging and reporting. Logging templates have been proposed to record information about data and models that might reveal risks or issues. Model Cards for Model Reporting provides open-ended questions to facilitate the recording of benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups [32]. Similarly, Datasheets for Datasets facilitates the logging of the data's characteristics, recommended uses, and other information [15]. While these logs and records are important and may direct the logger to unexpected risks, they are not targeted specifically at risk identification – additional steps are needed to understand the implications of each of these recorded aspects.

2.1.7 (7) Communication and consultation. There have been some studies on how the results of the risk analysis may be communicated, including discussions of accountability, approval processes, and stakeholder engagements [24, 45].

2.1.8 (8) Monitoring and review. Some technical methods allow for monitoring and automated tests for ML models that are deployed and operational in live environments to ensure they are performing as expected. This includes fairness toolkits [31, 48] that calculate and track fairness annd performance metrics across model iterations. Certain risks, e.g. representation bias in new data, may be monitored with relevant techniques, e.g. anomaly detection [44]. Some have produced fairness and ethics checklists [11, 29] drawing from other domains, e.g. pre-flight pilot checklists and pre-surgery medical checklists. Madaio fairness checklist is six pages long with prompts such as "Solicit input on concerns on system vision." The DEON checklist covers ethics with 21 prompts, of which six explicitly address fairness/bias.

While checklists may seem similar to a risk identification questionnaire, they play a different role in RMP. Checklists are best aligned to the review of risks (label 8 in Fig. 2) to ensure that the most crucial, known issues have been considered and addressed. A checklist's goal is *verification* [14], serving as memory aid to enhance task performance, preventing human errors, and supporting quality control [40]. The *Checklist Manifesto* recommends they fit in one page, as they provide "reminders of minimum necessary steps and make them explicit," intended as a rapid process rather than a prompt for extended discussion [14]. Among the studies on when checklists fail, one frequently cited feature is *ambiguity* [10, 14, 43, 46]. Sidebottom et al. claim literature on checklist design converges on defining the optimal features: concise, unambiguous, sharply defined so that each item is concrete, actionable, with a clearly identifiable start and finish, specific about what, when, how and who should do what item, and easy to follow and compatible with standard practice.

The risk identification process (Fig. 2, (2)) is not amenable to such a precise, unambiguous format, considering the complex socio-technical contexts. A checklist would require consensus on how to detect and assess unfair bias, still a contested topic in academia. Studies on checklists also emphasise the importance prioritising the most crucial tasks, but risk identification precedes any analysis of potential impact (Fig. 2, (4)). Prioritisation of risks at this phase is infeasible because the risks have not yet been fully identified. ISO 31000 RMP involves finding, recognising, and describing risks, identifying possible sources of risks, events and circumstances that may influence the achievement of objectives, possible causes of risks, and potential consequences [9]. Our questionnaire aims at providing systematic guidance to navigate this diagnostic process. As our case study will demonstrate, the questionnaire prompts the user to consider a comprehensive range of potential sources of unintended bias.

2.2 Gap in risk identification tools

While the proposed methods just discussed represented important steps towards operationalising some of these concepts and frameworks into methods, much work is yet to be done to integrate them into business processes. In particular, there is a gap in *2.* ***Risk identification***: while frameworks have recently been proposed for risk identification of unintended biases, there has not yet been methods to operationalise them. While Suresh and Guttag [41] provide two case studies of unintended biases, they do not provide any generalised method to identify them. This is a gap we focus on and address in our paper. Specifically, we develop an approach for risk identification for unintended bias in model development lifecycle. This is our focus area in the process because only once bias risk is identified can it be evaluated, quantified, and mitigated, and it represents a significant gap in implemented methods. This supplements the previously proposed methods in an end-to-end bias RMP.

Phase	Related work	Gaps
1. Scope, context, criteria	Contextual applications [17, 25, 37]	Practitioner struggle to adapt methods and framework to their specific use cases, e.g. chatbots
2. Risk identification	Bias type frameworks [30, 35, 41]	Frameworks introduced to categorise types of biases are not yet operationalised into a practical tool or method
3. Risk analysis	Fairness tests [7, 12, 16, 18, 23, 47], open source toolkits [2, 31, 48, 50]	Recent studies show gaps in open source tools in functionality and usability
4. Impact assessment	Trade-off analysis [22, 26], long-term impact [28], data protection impact assessments [20]	There are many competing impact assessment methods
5. Risk mitigation	"De-biasing" [2, 3, 13, 21, 50]	Non-technical mitigation strategies are not addressed, and practitioners struggle to identify the most appropriate mitigation strategy
6. Risk logging & reporting	Templates [15, 32]	Additional steps are needed to identify the risks from the logs
7. Communication & consultation	Accountability [45], business processes and stakeholder engagement [24],	There are limited studies on how risk may be communicated
8. Monitoring & review	Fairness tools to monitor metrics [31, 48], checklists [11, 29]	This stage presumes risks are already unambiguously identified when unintended bias is a complex area of research with limited consensus on how it should be identified, tracked, and mitigated

Table 1: Related work and ISO 31000 (indicative examples)

3 RISK IDENTIFICATION QUESTIONNAIRE FOR DETECTING UNINTENDED BIAS

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 [19, 27, 45]. We introduce a risk identification questionnaire that helps to detect the potential risks for each type of bias in each phase of ML development lifecycle. To our knowledge, there is no other questionnaire that operationalises the framework of bias types to systematically identify unintended biases, that covers issues arising from each stage of the ML lifecycle. The checklist by Madaio et al. [29] raises particular considerations for practitioners to consider, but it is not aligned to any bias type frameworks and describes *activities* rather than questions helping to elucidate bias risks. For example, the checklist items include "solicit inputs and concerns on system vision" and "undertake user testing" with some example considerations. In contrast, our questionnaire is not intended as a checklist, but rather, aims to *identify whether or not a type of bias exists*, e.g. "are any of the recorded features affected by human judgment" detects data measurement biases.

The RMPs in scope of this questionnaire are highlighted in Fig. 2: the initial risk identification (label 2), with some necessary scoping (label 1) to contextualise the potential risks. The questionnaire aims to help practitioners systematically uncover unexpected bias risks, which would be further assessed through subsequent phases of analysis, impact assessment, and mitigation. By understanding not only *how* the model may be biased but also *why* the bias exists through an explicit identification of the risk type, the questionnaire allows for a more targeted assessment of impact and design of a mitigation strategy.

Note the questionnaire is not intended to be a comprehensive, definitive standard for bias risk assessment. Rather, it seeks to be general, providing a starting point for extension and customisation to a particular domain or scenario. Future work could further adapt the questionnaire and develop additional guidance on how it may be applied in different contexts. The risk identification process may be carried out internally (through different organisational teams) 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. 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).

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 2: Questionnaire structure

As Table 2 shows, the structure aligns to the bias framework of Suresh and Guttag in Fig. 1. After establishing context, subsequent sections ask probing questions for each stage of the model development lifecycle. Answering "yes" indicates a risk of bias in that phase, prompting its analysis, impact assessment, and mitigation in the further RMP stages (Fig. 2). The full questionnaire can be found in supplementary materials and at https://github.com/michelleslee/bias_in_lifecycle. A small sample of the questionnaire is displayed in Fig. 3.

4 QUESTIONNAIRE APPLIED TO INSURANCE FRAUD

We will now walk through a case study to demonstrate the types of bias risks that are identified. This was based on an interview with the developer of a fraud prediction model for an insurance company. 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. We also provide a summary table in the Appendix.

4.1 Practitioner's answers to questionnaire

4.1.1 (A) Background information. The questionnaire begins by probing on the potential positive and negative impacts of the model. 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 as a deliberate bar to making claims. There is also potential representational harm, i.e. 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.

4.1.2 (B) Design: historical/external bias. This section 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. 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.

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 neighborhoods, 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

Figure 3: Sample snapshot of the questionnaire

The practitioner 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 it is 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.

4.1.3 (C) Data collection: representation bias. Collection methodologies can skew how the data set represents the ground truth. Based on the completed questionnaire, there were four representation biases identified. 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 population.

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.

4.1.4 (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 practitioner 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 be themselves be biased and affected by developer judgement. There is also a risk of proxy measurement: any attempts to locate geographical patterns of fraud could create unintended correlations with certain national or racial groups. 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.

4.1.5 (E) Model build and training: aggregation bias. In searching for potential biases in model build processes, the questionnaire attempts to uncover aggregation biases, i.e. when populations are heterogeneous in a way such that a single model cannot account for all subgroups. The practioner 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 is possibly improperly aggregating different types of fraud with different causal mechanisms.

4.1.6 (F) Model evaluation: evaluation bias. The questionnaire then considers whether the model is over-fitting to a particular metric, e.g. accuracy only. 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 mean a sub-optimal customer experience, but a false negative involves a financial loss to the company. Both metrics are considered. In answering this section, 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.

4.1.7 (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, e.g. inter-connected models or embedded in human processes. It was 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.

4.2 Mitigation strategies

The questionnaire (through Sections A-G) led to the practitioner identifying several disparate bias types. Section (A) identifies the context. Each subsequent section addresses one type of bias, facilitating the design of mitigation strategies appropriate for that bias type. We now 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 the previous section was done by the practitioner, this section represents our own response to the issues raised.

4.2.1 (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.

4.2.2 (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 [5].

4.2.3 (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 to racial or religious groups. The model could be trained on confirmed instances of fraud and on investigation results in addition to those correctly flagged.

4.2.4 (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.

4.2.5 (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.

4.2.6 (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.

4.3 Actionable insights

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. We now present an indicative set of action points following the risk identification that demonstrates the potential for this approach.

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 (Fig. 2 (label 4)). 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 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 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 data and models, it shows the potential for an informative impact assessment related to unintended biases. For example, Model D is the most accurate at identifying true fraud, but it also has one of the top false positive rates 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 instead. Model B is worse than F or G so can be removed, etc.

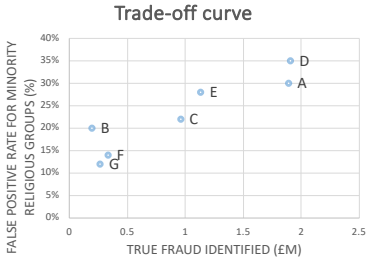

Figure 4: Example trade-offs in fraud detection model

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 (See Fig. 2), 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.

5 PRACTITIONER SURVEY

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, e.g. 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 who started the survey, 78 (74%) of the respondents completed at least one section and 29 (28%) completed the entire survey. The survey and its summary statistics, along with the full questionnaire, can be found in the supplementary materials.

Note that the questions 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.

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. 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) [4] to measure usability.

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.

5.1 Uncovering unexpected biases

Our results show that bias is clearly of concern. Our survey confirmed that 90% of practitioners believe the "ability to proactively diagnose unexpected issue(s)" is extremely/very important. 86% of them agree that our proposed questionnaire meets this need. Practitioners commented that the "breakdown of different types of biases," "clear structure," "standardizing 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.

5.2 Integration

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". Several answered it "can be integrated straight away" and "would fit in well with our existing risk management, documentation, and approval processes." A few who disagreed explained it was not directly relevant to their work, one stating it would require domain-specific modifications and adaptations.

5.3 Facilitating communication

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."

5.4 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 (See Fig. 2), 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 (Fig. 2 (5)). This will be further discussed in §6.

5.5 Usability

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)[1]. 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 [1], 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) [1]. 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. In the SUS survey, 59% of the interviewees agreed with: "I think I would like to use this questionnaire frequently." One said, "I would probably bookmark the questionnaire," signifying its potential for wider adoption. 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 our questionnaire design.

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 [16, 39]. Weighting each risk in a scoring system would also only be feasible once biases and their impact are understood in the further analysis stages (Fig. 2 (3-4)), which are out of scope for this paper.

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 et al. 2012) that correct model outputs to equalise a given metric. This misaligns with the survey'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. calls 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. debiasing) may fit in as part of a broader mitigation strategy, they should not be treated as a panacea for all bias risks.

5.6 Perceived relevance

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 because they do not use personal data. Two of the survey respondents also said there are no resources allocated on this issue because of limited business incentive or lack of awareness. However, 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 [33], 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, which should be further explored in future studies. That said, 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.

6 DISCUSSION AND FUTURE WORK

Our goal was to introduce a risk identification questionnaire to help practitioners identify potential bias risks. The survey shows practitioners find the questionnaire helpful, particularly in its *breakdown* of bias types introduced in past frameworks, in order to identify where 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.

Our findings reveal several opportunities for future research. The first area 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, the high uptake signals practitioner interest in ML bias issues. However, the practitioners' narrow understanding of model biases 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 addressing the practitioners' expressed needs and preferences. Our results highlighted key criteria expressed by the practitioners that are relevant more broadly to methods introduced in ML bias. For tools across bias RMPs, practitioners stressed the importance of alignment with domain-specific use cases, echoing past work [19, 27].

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 (Fig. 2), this presents an important challenge for future work, in particular because not all mitigation strategies are obvious, and the lack of consensus in literature. Suresh and Guttag 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." 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.

CONCLUSION

In this paper, 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 is 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 and linking it to a standard RMP. The practitioners surveyed were generally in agreement that the questionnaire met their requirements.

Our work represents but a first 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. Future work should address the other phases of the end-to-end RMP, including impact assessments and mitigations beyond the scope of our paper, to ensure the proposed methods are well-aligned to industry standards and easy to integrate to existing practices. This would help practitioners in implementing a more effective end-to-end bias risk management process.

ACKNOWLEDGEMENTS

We acknowledge the financial support of Aviva, the UK EPSRC (EP/P024394/1, EP/R033501/1) and Microsoft via the Microsoft Cloud Computing Research Centre.

REFERENCES

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

[2] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, A Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development 63, 4/5 (2019), 4–1. https://arxiv.org/abs/1810.01943

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. arXiv preprint arXiv:1706.02409 (2017).

[4] John Brooke. 1996. SUS: A quick and dirty usability scale. In Usability evaluation in industry. Taylor and Francis.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16 (2002), 321–357.

[6] Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2021. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Virtual Event, Canada) (FAccT '21). Association for Computing Machinery, New York, NY, USA, 598–609. https://doi.org/10.1145/3442188.3445921

[7] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 (2018).

[8] COSO. 2017. Committee of Sponsoring Organizations of the Treadway Commission Enterprise Risk Management Integrated Framework. https://www.coso.org/Pages/erm.aspx

[9] Alex Dali and Christopher Lajtha. 2012. ISO 31000 risk management—"The gold standard". EDPACS 45, 5 (2012), 1–8.

[10] Asaf Degani and Earl L Wiener. 1993. Cockpit checklists: Concepts, design, and use. Human factors 35, 2 (1993), 345–359.

[11] DrivenData. 2019. Deon: An ethics checklist for data scientists. DrivenData. http://deon.drivendata.org/

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

[13] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 259–268.

[14] Atul Gawande. 2010. Checklist manifesto, the (HB). Penguin Books India.

[15] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. arXiv preprint arXiv:1803.09010 (2018).

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

[17] Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In Proceedings of the 2018 World Wide Web Conference. 903–912.

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

[19] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–16.

[20] Margot E Kaminski and Gianclaudio Malgieri. 2020. Multi-layered explanations from algorithmic impact assessments in the GDPR. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 68–79.

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

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

[23] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. arXiv e-prints, Article arXiv:1703.06856 (March 2017), arXiv:1703.06856 pages. arXiv:1703.06856 [stat.ML]

[24] Michelle Lee, Luciano Floridi, and Alexander Denev. 2020. Innovating with Confidence: Embedding AI Governance and Fairness in a Financial Services Risk Management Framework. Berkeley Technology Law Journal 34, 2 (2020).

[25] Michelle Seng Ah Lee and Luciano Floridi. 2020. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. Minds and Machines (2020). https://doi.org/10.1007/s11023-020-09529-4

[26] Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. 2021. Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. Available at SSRN (2021). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3679975

[27] Michelle Seng Ah Lee and Jatinder Singh. 2021. The Landscape and Gaps in Open Source Fairness Toolkits. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) (CHI '21). Association for Computing Machinery, Yokohama, Japan, 13 pages. https://doi.org/10.1145/3411764.3445261

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

[29] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. 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 (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376445

[30] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019).

[31] Microsoft. 2019. Fairlearn. https://fairlearn.github.io/

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

[33] Sauradip Nag, Palaiahnakote Shivakumara, Yirui Wu, Umapada Pal, and Tong Lu. 2018. New COLD Feature Based Handwriting Analysis for Enthnicity/Nationality Identification. In 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE, 523–527.

[34] Arvind Narayanan. 2018. Tutorial: 21 Definitions of Fairness and their Politics. YouTube. https://www.youtube.com/watch?v=jIXIuYdnyyk

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

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

[37] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. 2018. Ensuring fairness in machine learning to advance health equity. Annals of internal medicine 169, 12 (2018), 866–872.

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

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

[40] Aiden Sidebottom, Nick Tilley, and John E Eck. 2012. Towards checklists to reduce common sources of problem-solving failure. Policing: A Journal of Policy and Practice 6, 2 (2012), 194–209.

[41] Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002 (2019).

[42] Hans Thamhain. 2013. Managing risks in complex projects. Project management journal 44, 2 (2013), 20–35.

[43] Jonathan R Treadwell, Scott Lucas, and Amy Y Tsou. 2014. Surgical checklists: a systematic review of impacts and implementation. BMJ quality & safety 23, 4 (2014), 299–318.

[44] Franco van Wyk, Yiyang Wang, Anahita Khojandi, and Neda Masoud. 2019. Real-time sensor anomaly detection and identification in automated vehicles. IEEE Transactions on Intelligent Transportation Systems 21, 3 (2019), 1264–1276.

[45] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. 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. 1–14.

[46] EGG Verdaasdonk, LPS Stassen, Prama P Widhiasmara, and Jenny Dankelman. 2009. Requirements for the design and implementation of checklists for surgical processes. Surgical endoscopy 23, 4 (2009), 715–726.

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

[48] Matthijs Vincent. 2019. scikit-fairness. https://github.com/koaning/scikit-fairness

[49] Robert James Voetsch, Denis F Cioffi, and Frank T Anbari. 2004. Project risk management practices and their association with reported project success. In Proceedings of 6th IRNOP Project Research Conference, Turku, Finland. Citeseer, 680–97.

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

[51] Ofer Zwikael and Mark Ahn. 2011. The effectiveness of risk management: an analysis of project risk planning across industries and countries. Risk Analysis: An International Journal 31, 1 (2011), 25–37.