NUS Law Working Paper No 2022/007

A Legal Framework for Artificial Intelligence Fairness Reporting

Yap Jia Qing
Ernest Lim

ernestlim@nus.edu.sg

[June 2022]

(2022) 81 *Cambridge Law Journal* (forthcoming)

A LEGAL FRAMEWORK FOR ARTIFICIAL INTELLIGENCE FAIRNESS REPORTING

Yap Jia Qing* and Ernest Lim**

ABSTRACT. *Clear understanding of artificial intelligence (AI) usage risks and how they are being addressed is needed, which require proper and adequate corporate disclosure. We advance a legal framework for AI Fairness Reporting to which companies can and should adhere on a comply or explain basis. We analyse the sources of unfairness arising from different aspects of AI models and the disparities in the performance of machine learning systems. We evaluate how the machine learning literature has sought to address the problem of unfairness through the use of different fairness metrics. We then put forward a nuanced and viable framework for AI Fairness Reporting comprising: (a) disclosure of all machine learning models usage; (b) disclosure of fairness metrics used and the ensuing trade-offs; (c) disclosure of de-biasing methods used; and (d) release of datasets for public inspection or for third-party audit. We then apply this reporting framework to two case studies.*

KEYWORDS: artificial intelligence, machine learning, fairness, discrimination, disclosure, reporting, companies, law and technology, shareholders, stakeholders, GDPR.

I. INTRODUCTION

Regulatory bodies and think tanks across the world have published reports and guidelines on the ethical use of artificial intelligence (AI), but generally hesitate to take a command-and-control approach to AI regulation coupled with the imposition of sanctions due to the rapidly evolving nature of AI and the lack of clarity even within the technical community on how ethical ideals can be operationalised.[1]

*Visiting Researcher, Centre for Technology, Robotics, Artificial Intelligence and the Law, National University of Singapore.

**Professor, Faculty of Law, National University of Singapore. Address for Correspondence: Faculty of Law, National University of Singapore, 469G Bukit Timah Road, Singapore 259776. Email: lawlimw@nus.edu.sg; ttycd_t@yahoo.com. We are grateful to Simon Chesterman and the two Cambridge Law Journal anonymous referees for their insightful comments. The usual disclaimers apply.

[1] See eg, OECD Principles on AI (2019); State of implementation of the OECD AI Principles: Insights from National AI Policies (OECD Digital Economy Papers, June 2021, No. 311).

Other than command and control regulation on the fairness of AI use[2] (which has been said to stifle innovation[3]), a less intrusive approach could be through reflexive regulation in the form of AI Fairness Reporting, similar to sustainability/environmental, social and governance (ESG) reporting.[4] The risks from the unfair provision and use of AI systems have already made their way into mainstream financial filings as a material risk, with Microsoft's 2021 Annual Report warning that: "AI algorithms may be flawed. Datasets may be insufficient or contain biased information … If we enable or offer AI solutions that are controversial because of their impact on human rights, privacy, employment, or other social issues, we may experience brand or reputational harm".[5]

There are well-mapped legal risks, regulatory risks, reputational risks and the risk of financial and operational losses from the use of AI.[6] General statements about AI risk as seen in Microsoft's annual report are not sufficient for shareholders and other stakeholders to assess the full extent of fairness risks faced by the company in the provision and use of AI. Besides, increased investor awareness of sustainable investing would lead to investors demanding to know whether artificial intelligence solutions used or sold by companies are aligned with their values.

AI Fairness Reporting beyond general statements relating to AI risks in annual reports or other filings would require standards akin to the Global Reporting Initiative (GRI) standards in sustainability reporting.[7] Similar to how sustainability reporting rules (and practice notes) require (or advise) companies to describe both the reasons and the process of selecting material ESG factors[8], there is a need to work towards specifying the substantive content of what AI fairness metrics can be reported, in a manner which will be useful for public scrutiny and debate by stakeholders, regulators and civil society.

---

[2] See the European Commission "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts" (COM/2021/206 final). The draft regulations impose obligations on providers and users of AI systems by distinguishing three types of risks. Providers of high risk AI systems are required to put in place risk management systems, technical documentation, quality management system, and conformity assessment processes.

[3] A. Mcafee, "EU Proposals to Regulate AI Are Only Going to Hinder Innovation" *Financial Times* (26 July 2021).

[4] For the industry benchmark, see the Global Reporting Initiative Standards, available at https://www.globalreporting.org/how-to-use-the-gri-standards/ (last accessed 13 November 2021).

[5] Microsoft, Form 10-K (Annual Report for the fiscal year ended June 30, 2021), 28. It is striking that one of the world's top three largest and most influential technology companies merely devotes fewer than ten sentences to the risks of AI in its more than 100-page annual report.

[6] See eg, I. Chiu and E. Lim, "Managing Corporations' Risk in Adopting Artificial Intelligence: A Corporate Responsibility Paradigm" (2021) 19 Washington Univ. Global Studies L. Rev. 347. The unique regulatory challenges arising from the opacity of AI are comprehensively explored in S. Chesterman, "Through a Glass, Darkly: Artificial Intelligence and the Problem of Opacity" (2021) 69 A.J.C.L. 271. Companies have to balance these risks with the massive potential gains from use of AI, which come in the form of both revenue increase and cost reduction. Revenue increase is associated with AI adoption in pricing and promotion, inventory and parts optimization, customer service analytics, as well as sales and demand forecasting. Cost reduction results from optimisation of talent management, contact centre automation, and warehouse optimisation. See McKinsey Analytics, "The State of AI in 2020" (November 17, 2020).

[7] GRI, available at https://www.globalreporting.org/how-to-use-the-gri-standards/ (last accessed 13 November 2021).

[8] See eg, Regulation (EU) 2019/2088 of the European Parliament and of the Council of 27 November 2019 on sustainability-related disclosures in the financial services sector; Proposal for a Directive of the European Parliament and of the Council, amending Directive 2013/34/EU, Directive 2004/109/EC, Directive 2006/43/EC and Regulation (EU) No 537/2014, as regards corporate sustainability reporting COM/2021/189 final.

Unfortunately, current guidance on Data Protection Impact Assessments (DPIA) under the General Data Protection Regulation (GDPR) do not make reference to the development of metrics which capture different notions of fairness in the technical machine learning literature.[9] In this paper, we propose a legal framework for AI Fairness Reporting informed by recent developments in the computer science machine learning literature on fairness. Companies should disclose the fairness of machine learning models produced or used by them based on our proposed framework on a comply or explain basis.[10]

The argument for a framework for AI Fairness Reporting comprises five parts. First, reasons are given as to why a reporting framework is needed. Second, the nature or sources of unfairness is identified. Third, how the machine learning literature has sought to address the problem of unfairness through the use of fairness metrics is analysed. Fourth, bearing in mind the issues related to unfairness and the fairness metrics, we propose a legal solution, namely, what the disclosure contents of the AI Fairness Reporting framework should consist of. Fifth and finally, the proposed Reporting framework is applied to two case studies.

The structure of this article is as follows. Part II provides three reasons for having the AI Fairness Reporting Framework: (a) to enable a better understanding of the potential legal liability risks due to contravention of applicable legislation; (b) to address investors' and stakeholders' pro-social expectations concerning the company's business and operations; and (c) to address inadequacies in the DPIA under the GDPR.

Part III analyses the nature or sources of unfairness. The unfairness can arise from different aspects in the process of building a supervised machine learning model, specifically with regards to data creation and labelling as well as feature extraction, embeddings and representation learning. The unfairness can also arise from the disparities in the performance of machine learning systems with respect to data related to different demographic groups.

Part IV examines how the machine learning literature has sought to address the problem of unfairness by using different metrics of fairness. These metrics are analysed, followed by an assessment of the trade-offs between the fairness metrics and the disparities in AI model performance.

Part V advances a framework for AI Fairness Reporting, the proposed reporting obligations of which should include: (a) disclosure of all uses of machine learning models; (b) disclosure of the

---

[9] A. Kasirzadeh and D. Clifford, "Fairness and Data Protection Impact Assessments" 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; M.E. Kaminski et. al., "Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations" (2021) 11 International Data Privacy Law 125. Other than algorithmic impact assessment, other suggested solutions include AI Ombudsperson and AI audits, see S. Chesterman, *We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law* (Cambridge 2021), 154-7.

[10] Under a "comply or explain" mechanism, companies are required to comply with the rules, but can provide an explanation instead, if they choose not to comply. A comply or explain mechanism allows shareholders, stakeholders and the market to decide whether the explanation given by the company for not complying is satisfactory and, if not, to take action. See eg, I. MacNeil and I. Esser, "The Emergence of 'Comply or Explain' as a Global Model for Corporate Governance Codes" (2022) 33 Eur. Bus. L. Rev. 1.

fairness metrics used and the ensuing trade-offs; (c) disclosure of the de-biasing methods used; and (d) release of datasets for public inspection or for third-party audit.

Part VI applies the proposed AI Fairness Reporting framework to two case studies—one relating to credit profiling and the other to facial recognition—in order to show its utility. This is followed by the conclusion.

II. WHY THE NEED FOR AI FAIRNESS REPORTING

A. To Enable Stakeholders to Better Understand Potential Legal Liability Risks

A first practical reason for the need for AI Fairness Reporting is to empower stakeholders like investors, customers and employees of a company to better assess the legal risks of a company due to potential breaches of applicable legislation through its use of machine learning models. We consider statutory examples from the UK and the US.

1. Equality Act 2010

The forms of discrimination under the UK Equality Act can be divided into direct discrimination and indirect discrimination. Section 13(1) of the Equality Act defines direct discrimination as Person A treating Person B less favourably than Person A treats or would treat others, because of a "protected characteristic" of B. Section 14 of the Act sets out the concept of combined discrimination, where direct discrimination happens on the basis of two relevant protected characteristics. The protected characteristics include age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex and sexual orientation.[11]

Indirect discrimination under the UK Equality Act, as defined in Section 19, refers to the application of a provision, criterion or practice that puts people with a relevant protected characteristic at a 'particular disadvantage', without showing the provision, criterion or practice to be a proportionate means of achieving a legitimate aim. The difference from direct discrimination is that the provision, criterion or practice only needs to be related to the protected characteristic, and use of the protected characteristic itself is not needed for indirect discrimination to be found. For example, an algorithm used by a bank in relation to credit card applications that does not assign different credit worthiness based on the protected characteristics, but on spending patterns related to certain products and services, may impose a particular disadvantage on certain segments of the population, thus potentially violating the Equality Act.[12]

[11] See Equality Act 2010, s 4.

[12] To prove the indirect discrimination, the claimant must show the disadvantage as compared with a similarly situated individual (also known as a hypothetical comparator) who does not share the protected characteristic – this can be understood as counterfactual fairness as seen in the machine learning literature. The mere fact of a disadvantage, without the need for explanation from the claimant on why it occurs, puts the burden on the party which applies the provision, criterion or practice to justify it. In the UK, statistical evidence can be used to demonstrate the "particular

2. GDPR

The GDPR became a part of UK domestic law, in accordance with Section 3 of the European Withdrawal Act 2018. The GDPR governs the processing of personal data, and 'profiling' is defined under the GDPR as "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person".[13] Thus, most machine learning models acting on individuals will fall under this definition of profiling under the GDPR. Article 5 of the GDPR states the principle that data shall be processed 'lawfully, fairly and in a transparent manner', and GDPR Article 24(1) requires that 'appropriate technical and organisation measures' need to be implemented in light of risks to the rights of individuals.

Processing of 'special category data'[14] is prohibited under Article 9(1) of the GDPR, unless one of the exceptions in Paragraph 2 is satisfied. This concept of special category data is similar to that of protected characteristics discussed above regarding the UK Equality Act. However, this also means that a machine learning engineer is prevented from using special category data in the algorithm in order to correct for human biases in the dataset[15] unless the engineer fulfils one of the Paragraph 2 exceptions like consent. However, it has been argued that genuinely free consent cannot be obtained in this case, because a refusal to grant consent results in the individual suffering a higher risk of discrimination.[16]

In addition, it is unclear when multiple proxies available in the data allow for a 'special category' information to be inferred, such that the proxies are taken together to make up special category data. The UK's Information Commissioner's Office guidelines on special category data states that it depends on the certainty of the inference and whether the inference was deliberately drawn.[17] Courts, in interpreting this provision, will likely distinguish between (i) explicit inference of special category data through an output or intermediate step of an algorithm, and (ii) algorithms which produce outputs correlated with special categories.[18] Adding to the latter case, we think algorithms which have inputs correlated with special categories would belong to that category too, and this latter case should not trigger Article 9.

---

disadvantage", though no statistical threshold is set to delineate the permissible level of disadvantage, unlike in the US where the four-fifths rule is used by the Equal Employment Opportunity Commission. See A. Kelly-Lyth, "Challenging Biased Hiring Algorithms (2021) 41 O.J.L.S. 899.

[13] These aspects cover "… the natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements": see Article 4(4) of the GDPR.

[14] GDPR Article 9(1): "personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation".

[15] Kelly-Lyth, "Challenging Biased Hiring Algorithms"; J. Kleinberg et al, "Algorithmic Fairness" (2018) 108 AEA Papers and Proceedings 22; T.B. Gillis and J. Spiess, "Big Data and Discrimination" (2019) 86 Univ. Chicago L. Rev. 459.

[16] See Kelly-Lyth, "Challenging Biased Hiring Algorithms" who refers to GDPR recital 42. See also European Data Protection Board, Guidelines 05/2020 on consent under Regulation 2016/679, version 1.1. (4 May 2020) [13] which states that consent will not be valid if the data subject will endure negative consequences if they do not consent.

[17] ICO, "Special Category Data: What is Special Category Data?", available at https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/special-category-data/what-is-special-category-data/ (last accessed 13 November 2021).

[18] Kelly-Lyth, "Challenging Biased Hiring Algorithms".

3. Domain-specific regulations in the US

The US has domain-specific regulations in a variety of areas where machine learning is now applied, listing protected characteristics similar to the UK Equality Act. For example, the Fair Housing Act[19] in the United States lists race, colour, national origin, religion, sex, familial status and disability as protected characteristics. The United States' Equal Credit Opportunity Act also lists all of the above protected attributes with the exception of familial status and disability, but with the inclusion of exercised rights under the Consumer Credit Protection Act, marital status, status as a recipient of public assistance and age.

Employment law in the US also allows an employer to be sued under Title VII for employment discrimination under one of two theories of liability: disparate treatment and disparate impact.[20] Disparate treatment comprises either formal disparate treatment of similarly situated people or the intent to discriminate. Disparate impact refers to practices that are superficially neutral but have a disproportionately adverse impact on groups with protected characteristic. Disparate impact is not concerned with intent, but instead first asks whether there is a disparate impact on members of a group with a protected attribute, secondly whether there is a business justification for that impact, and finally, if there are less discriminatory ways of achieving the same result.[21] The US Equal Employment Opportunity Commission advocates for a four-fifths rule[22] – the ratio of the probability of one group of the protected characteristic getting hired, over the probability of the other group with the protected characteristic getting hired, should not be lower than four-fifths.

AI Fairness reporting would allow stakeholders and regulators to better assess whether sufficient work has been done by the company to be compliant with such regulations. Reporting on fairness of AI models would help to assure stakeholders about the reputational risks of the company being involved in a discrimination scandal, especially when such incidents can impact share prices and result in the loss of talent.

B. Sustainable Investments

There has been a rapid growth in sustainable investments in the last several years, which incorporate various ESG related concerns or objectives into investment decisions. Globally, assets under management in ESG mutual funds and exchange-traded funds has grown from $453 billion in 2013 to $760 billion in 2018, and is expected to continue growing.[23] It is plausible that AI fairness considerations are already being taken into account by such ESG funds, if not in the near future, in assessments under the governance pillar of ESG. There is already work from investment

[19] Sec. 804 42 U.S.C. 3604.
[20] S. Barocas and A.D. Selbst, "Big Data's Disparate Impact" (2016) 104 Cali. L. Rev. 671.
[21] Ibid.
[22] The U.S. EEOC. Uniform Guidelines on Employee Selection Procedures (March 2, 1979).
[23] BlackRock, "Sustainability: The Future of Investing" (February 2019).

funds on establishing a set of requirements including non-bias and transparency of AI use, as a yardstick by which companies can be evaluated for corporate stewardship where AI is applied.[24]

Stakeholder capitalism, which challenges the idea of shareholder primacy, seeks to promote long-term value creation by taking into account the interests of all relevant stakeholders.[25] Stakeholder capitalism is premised on the idea that the stock market misvalues intangibles that affect stakeholders, like employee satisfaction.[26] Therefore, it emphasizes that corporate directors and executives should make decisions in a manner which takes into account the interests of other stakeholders like customers, employees and society at large. A natural extension of the considerations which stakeholder capitalism would have corporate directors and executives take into account would be whether AI products and services used or sold by the company are fair towards potential job applicants, employees, customers and individuals of the public.

C. Inadequacies in the DPIA under the GDPR

The GDPR requires that a DPIA be carried out for any data processing which is 'likely to result in a high risk to the rights and freedoms of natural persons'.[27] This reference to the 'rights and freedoms of natural persons' is interpreted to be concerned not only with the rights to data protection and privacy, but also, according to the Article 29 Data Protection Working Party Statement on the role of a risk-based approach in data protection legal frameworks, with other fundamental rights including the prohibition of discrimination.[28] Examples of processing operations which are "likely to result in high risks" are laid out in Article 35(3). Article 35(3)(a) relates to "a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person". This is further elaborated on in recital 71 which specifies processing operations like "analysing or predicting aspects concerning performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, in order to create or use personal profiles". Further, Article 35(3)(b) relates to "processing on a large scale of special categories of data referred to in Article 9(1), or of personal data relating to criminal convictions and offences referred to in Article 10". Recital 75 explains such special categories of data as those which "reveal racial or ethnic origin, political opinions, religion or philosophical beliefs, trade union membership and the processing of genetic data, data concerning health or data concerning sex life or criminal convictions and offences or related security measures".

---

[24] "Investors' Expectations on Responsible Artificial Intelligence and Data Governance", Hermes Investment Management (April 2019).
[25] K. Schwab and P. Vanham, *Stakeholder Capitalism: A Global Economy that Works for Progress, People and Planet* (Wiley, 2021); McKinsey and Company, *The Case for Stakeholder Capitalism* (12 November 2020).
[26] Ibid.
[27] GDPR art 35.
[28] Article 29, Data Protection Working Party, 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 (17/EN WP 248 rev.01), 6.

However, the exact scope and nature of what a DPIA entails, especially relating to issues concerning fairness, is less clear. Article 35(7) of the GDPR, read with Recital 84 and 90, sets out the minimum features of a DPIA to comprise "a description of the envisaged processing operations and the purposes of the processing", "an assessment of the necessity and proportionality of the processing", "an assessment of the risks to the rights and freedoms of data subjects", and the measures envisaged to "address the risks" and "demonstrate compliance with this Regulation".[29] The methodology of the DPIA is left up to the data controller. Even though guideline criteria are provided[30], they make no reference to any fairness metrics and debiasing techniques which have emerged in the technical machine learning literature[31].

Although prior work on biased hiring algorithms called for DPIA reports to be made available publicly[32], there is no current requirement under the GDPR for such DPIA reports to be made public. Moreover, we do not think DPIA reports in its current form as defined under the GDPR and its guidance documents adequately serve the needs of AI Fairness Reporting because the DPIAs do not require the disclosure of fairness metrics and the debiasing methods used.[33]

III. SOURCES OF UNFAIRNESS IN THE MACHINE LEARNING MODELS AND PERFORMANCE

A. Unfairness from the Process of Building Supervised Learning Models

We first examine how bias can be attributed to the various stages of the process of building supervised learning models. In general, there are three broad types[34] of machine learning models – supervised learning, unsupervised learning and reinforcement learning. Supervised learning models are trained on data examples labelled with the decision which should be made. These labels are created either by manual human labelling, or by less precise proxy sources or heuristics in a method known as weak supervision. When supervised models are trained using the labelled examples, the model learns how much weight to put on various factors fed to it when making a decision. In unsupervised learning, the data examples given to the model are not labelled with the decision– the model's goal here is simply to find patterns in the data, without being told what patterns to look for and with no obvious measure of how well it is performing. Distinct from supervised and unsupervised learning models, reinforcement learning models harness reward or punishment signals to learn how to act or behave. In our discussion, we focus primarily on supervised learning which has thus far brought about the most legal and policy concerns surrounding fairness.

[29] Ibid.
[30] Ibid, "Annex 2 – Criteria for an acceptable DPIA".
[31] Kasirzadeh and Clifford, "Fairness and Data Protection Impact Assessments".
[32] Kelly-Lyth, "Challenging Biased Hiring Algorithms".
[33] See Part V for an analysis of the debiasing methods.
[34] K.P. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT 2012).

1. Dataset creation and labelling

In the dataset creation process, unfair sampling can occur from operational practices in the company. A practice of refusing credit to minorities without first assessing them would result in records of minorities being less represented in the training data set[35]. By only observing the training data, it can be hard to detect the presence of a sample selection bias. Supervised learning models are dependent on the labels given to data in the training set. If the organisation had been making unfair decisions reflected in the training dataset, such unfairness will result in the trained model. For example, human essay graders are known to have prejudices on the linguistic choices of students which signify membership in demographic groups.[36] Automatic essay grading models might then be trained on a dataset of essays with the corresponding scores assigned by such human essay graders, thus learning the biases of the humans into the models.

2. Feature extraction, embeddings and representation learning

Although images and text are easily associated with meaning when presented to a human, in the raw form these data types are devoid of meaning to a computer. Raw images are just rows of pixel values, while text is just a string of characters each encoded in the ASCII[37] format. Deep neural network models are used to learn feature maps of images and embeddings of text which are used respectively in the computer vision and natural language processing applications of AI. For example, words can be represented in the form of vector embeddings, which can capture meaning and semantic relationships between words through their relationship with vectors representing other words. In the classic word2vec example, the direction and distance between the vectors representing the words king and queen, are similar to that of the direction and distance between the vectors representing the words husband and wife.

Traditionally, heuristics or rule-based approaches are used to create such features from the input data. Today, deep learning methods often rely on representation learning to learn the features or representations, by training on large datasets like CommonCrawl, using the sequential co-occurrence of words which appear close together in texts across the web as a signal to encode their semantic meaning. The principle behind how these embeddings are learnt is that "a word is characterized by the company it keeps"[38]. There is much technical evidence to show that contextualised word embeddings, which are often used as inputs to current state-of-the-art natural language processing systems, encapsulate gender biases.[39] An extensive study[40] looked into how

[35] T. Kamishima et. al., "Fairness-aware Classifier with Prejudice Remover Regularizer" 2012 Joint European Conference on Machine Learning and Knowledge Discovery in Databases.
[36] S. Barocas et al, Fairness and Machine Learning (Fairmlbook.org, 2019). See also R.N. Hanna and L.L. Linden, "Discrimination in Grading" (2012) 4 *Amn. Econ. J.* 146.
[37] American Standard Code for Information Interchange.
[38] An idea which originated in the field of distributional semantics in computational linguistics. See J.R. Firth, "A Synopsis of Linguistic Theory 1930–1955" (1957) Studies in Linguistic Analysis 1. Later used in Y. Bengio et al., "A Neural Probabilistic Language Model" (2013) 3 J. Machine Learning Research 1137 and in more recent models like M. Tomas et. al., "Distributed Representations of Words and Phrases and their Compositionality" (2013) Advances in Neural Information Processing Systems.
[39] C. Basta et. al., "Extensive Study on the Underlying Gender Bias in Contextualized Word Embeddings" 33 (2021) Neural Computing and Applications 3371.
[40] Ibid.

stereotypical associations between gender and professional occupations propagate from the text used to train the models to the text embeddings. The study covered four different domains – the medical domain using PubMed data, the political domain using Europarl data, a social domain using TEDx data and the news domain – and found such propagation of gender bias across domains. However, it is notable that less bias was found in the social domain which used TEDx data, and very evidently found in Pubmed which frequently associates medical occupations with male gender pronouns.

In the use of deep neural networks for supervised learning, engineers might face the practical problem of having insufficient labelled data in their dataset. This is especially so in applications where it takes domain experts to label the data, such that the creation of a huge, labelled dataset is a costly endeavour. To overcome the problem of limited training data, machine learning engineers would often use a technique named transfer learning. This technique involves using a model already trained on another (possibly larger) dataset similar to that of the data the engineer is working with, before continuing training on the limited labelled data. Open-source models pretrained on open datasets are made widely available by universities and technology companies. However, the geographic distribution of images in the popular ImageNet dataset reveals that 53% of the images were collected in the US and Great Britain, and a similar skew is also found in other popular open-source image datasets like Open Images.[41] This can lead to models trained on such datasets performing better in the recognition of objects more commonly found in the US and UK, than in other countries.

B. Unfairness through Disparities in the Performance of Machine Learning Models

Beyond the fairness of classification decisions produced by supervised learning models, a notion of fairness more generally applicable to all machine learning models that might not be clearly addressed by existing laws—but considered in the machine learning literature on fairness—is the disparities in the performance of machine learning models with respect to data related to different demographic groups. This can occur, for instance, when these groups are underrepresented in datasets used for training machine learning models. In addition, other applications of machine learning beyond classification can propagate bias when they are trained on datasets which are labelled by biased humans or biased proxy data.

1. Natural language processing

There are disparities between how well machine learning systems which deal with natural language perform for data relating to different demographic groups. Speech-to-text tools do not perform as well for individuals with some accents.[42] Sentiment analysis tools have been shown to systematically assign different scores to text based on race-related or gender-related names of

[41] N. Meharbi et. al., "A Survey on Bias and Fairness in Machine Learning" (2019) arXiv preprint arXiv:1908.09635.
[42] R. Tatman, "Gender and Dialect Bias in YouTube's Automatic Captions," in First ACL Workshop on Ethics in Natural Language 2017, 53-59.

people mentioned[43], while annotators' insensitivity to differences in dialect have also resulted in automatic hate speech detection models displaying a racial bias, such that words and phrases which are characteristic of African American English are correlated with ratings of toxicity in numerous widely-used hate speech datasets, which were then acquired and propagated by models trained on these datasets. [44] Even relative to human graders who may themselves give biased ratings, automated essay grading systems tend to assign lower scores to some demographic groups in a systemic manner.[45]

It was found that when the sentences "She is a doctor. He is a nurse." were translated using Google Translate from English to Turkish and then back to English, gender stereotypes were injected, such that Google Translate returned the sentences "He is a doctor. She is a nurse."[46] The explanation provided by the researchers in the study is that Turkish has gender neutral pronouns, so the original gender information was lost during the translation from English to Turkish, and when the sentences were translated from Turkish back to English, the Google Translate picked the English pronouns which best matched the statistics of the text it was trained on.

2. Computer vision

Machine learning is widely deployed in computer vision tasks like image classification, object detection and facial recognition. However, as previously discussed[47], populations outside the US and UK are underrepresented in the standard datasets used for training such models. These datasets, curated pre-dominantly by white, male researchers, reflect the world view of its creators.

Images of household objects from lower-income countries are significantly less accurately classified than those from higher-income countries.[48] The commercial tools by Microsoft, Face++ and IBM designed for gender classification of facial images were shown to perform better on male faces than female faces, with up to a 20.6% difference in error rate.[49] The classifiers were also shown to perform better on lighter faces than darker faces, and worst on darker female faces.

3. Recommendation systems and search

Recommendation and search systems control the content or items which are exposed to users, and thus bring about a unique set of fairness concerns[50]. First, the informational needs of some

[43] S. Kiritchenko and S.M. Mohammad, "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems," (2018) arXiv Preprint arXiv:1805.04508.
[44] M. Sap et. al., "The Risk of Racial Bias in Hate Speech Detection," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, 1668-78.
[45] C. Ramineni and D. Williamson, "Understanding Mean Score Differences Between the e-rater Automated Scoring Engine and Humans for Demographically Based Groups in the GRE General Test," ETS Research Report Series (2018) 1-31.
[46] S Barocas et al, *Fairness and Machine Learning* (Fairmlbook.org 2019).
[47] Meharbi, "A Survey on Bias and Fairness in Machine Learning".
[48] T. de Vries et. al., "Does Object Recognition Work for Everyone?" in 2019 ICCV 52-59.
[49] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in 2018 FAccT, 77-91.
[50] S Barocas et al, *Fairness and Machine Learning* (Fairmlbook.org 2019).

searchers or users may be served better than those of others. Harms to consumers can happen when a recommendation system underperforms for minority groups in recommending content or products they like. Such unfairness is difficult to study in real systems as the relevant target variable of satisfaction is hard to measure[51] – clicks and ratings only serve as crude proxies to user satisfaction. Second, inequities may be created between content creators or product providers by privileging certain content over others. Youtube was sued in 2019 by content creators who alleged that the reach of their LGBT-focused videos was supressed by Youtube algorithms, while allegations relating to search have included partisan bias in search results.[52] Third, representational harms can occur by the amplification and propagation of cultural stereotypes.

4. Risk assessment tools

In risk assessment tools like COMPAS[53], calibration[54] is an important goal. Equalised calibration requires that 'outcomes are independent of protected attributes after controlling for estimated risk'.[55] For example, in a group of loan applicants estimated to have a 20% chance of default, calibration would require that the rate of default of whites and African Americans is similar, or even equal if equalised calibration is enforced. If a tool for evaluating recidivism risk does not have equalised calibration between demographic groups defined by race, the same probability estimate given by the tool would have a different meaning for African American and white defendants – inducing judges to take race into account when interpreting the predictions of the risk tool.[56]

IV. COMPETING ALGORITHMIC FAIRNESS METRICS AND TRADE-OFFS

A. Fairness Metrics of Supervised Classification Models

Although the concept of fairness[57] in the law governing data processing is nebulous, the technical machine learning community has developed several technical metrics of fairness. In this section, we attempt to give a flavour of the various main categories of technical fairness metrics.

---

[51] Ibid.
[52] Ibid.
[53] Correctional Offender Management Profiling for Alternative Sanctions. It assesses the risk of recidivism for offenders.
[54] Intuitively, calibration means that the probability estimate (confidence level of the decision) given by the model for its decisions carries semantic meaning. If the probability estimate given by the model for a set of 100 people in the dataset is 0.6, it means that 60 out of the 100 people should belong to the positive class. See G. Pleiss et. al., "On Fairness and Calibration" 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
[55] S. Corbett-Davies and S. Goel, "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning" (2021) 33 Neural Computing and Applications 3371.
[56] Pleiss, "On Fairness and Calibration".
[57] In data protection, fairness in the context of fair data use was initially understood as transparency. This understanding is rooted in the relation drawn between fairness and the expectations of the data subject, such that the data processor would have to accessibly inform data subjects about how and why personal data is processed. Such

To begin with, "Fairness through Unawareness" is an approach to machine learning fairness where the model simply ignores protected characteristics. This approach has been shown to be ineffective because it is possible for the model to infer information about such protected characteristics from other features, termed proxies or redundant encodings which are correlated with the protected characteristic[58], thus leading to indirect discrimination. A classic example of this would be the case of removing the protected attribute of race in a data set, but keeping another feature of the dataset on whether or not the individual visits the Mexican market on a weekly basis, which is correlated with the Hispanic race. Fairness through Unawareness, apart from being ineffective, requires all protected characteristics to be masked out. This requirement might be infeasible in some applications where it would, for example, require the removal of gender from facial images, or the removal of words relating to protected attributes from sentences which would be left devoid of readability.

To address the problems of Fairness through Unawareness, at least four fairness metrics have been developed which do without the need to mask out protected characteristics, and instead determine fairness directly based on the protected characteristic.[59] These four metrics are "Demographic Parity", "Equality of Odds", "Equality of Opportunity" and "Equalised Calibration". These metrics are examined in the context of a binary classification model, which is a machine learning model which outputs either a positive or negative class (for example, whether a person is positive or negative for a disease).

1. *Demographic parity*

transparency requirements are well-addressed in data protection regulations. Subsequent guidance from the UK Information Commissioner's Office expanded that understanding of fair data use to include the requirement of not using the data in a manner which will have "unjustified adverse effects", thus bringing it closer to the technical metrics of fairness examined in this article which are largely centered around having equal outcomes for different demographic groups: see "Data: A New Direction", Department for Digital, Culture, Media and Sport (21 September 2021), 26-31. To be clear, we are not endorsing the view that the law should accept the fairness metrics examined in this article as definitive to the exclusion of other notions of fairness (such as procedural fairness and loss of agency, which are not as easily measurable in the form of technical metrics and are thus better addressed through regulations which limit or proscribe certain behavior). After all, no prescriptive rules and no sanctions for violating such rules are proposed in this article. Instead, we argue that important information should be brought to light through disclosure along the lines of such fairness metrics which more adequately capture trade-offs or omissions that the company made in their design or use of AI. Armed with this important disclosure, shareholders and stakeholders can adopt whatever fairness standards or metrics they think apt. But they cannot make an informed choice without such a disclosure from the company because these fairness metrics reveal information about trade-offs and performance of the model which might not be that apparent without their disclosure.

[58] D Pedreshi, et al., "Discrimination-aware Data Mining*"* in Proc. 14th ACM SIGKDD, 2008.

[59] These four fairness metrics are group fairness metrics which make comparisons between demographic groups. Another approach to fairness is individual fairness which looks at whether similar individuals in the dataset are treated similarly. The technical metrics developed to further this approach measure the similarity between individuals. However, these similarity measures are often developed with feedback from human arbiters who might bring in their implicit or systemic biases, and the choice of fairness-relevant features to be used for evaluating similarity of the individuals is also morally laden. See W. Fleisher, "What's Fair About Individual Fairness?" (April 5, 2021), available at https://ssrn.com/abstract=3819799 (last accessed 13 November 2021).

The fairness metric of Demographic Parity measures how much an algorithmic decision is independent of the protected characteristic by taking the difference in the probability of the model predicting the positive class across demographic groups on the protected characteristic.[60] Between two demographic groups on the race protected characteristic, namely whites and African Americans, perfect satisfaction of this metric on a hiring model would result in the positive hiring decision being assigned to the two demographic groups at an equal rate.

However, there have been disadvantages[61] identified with Demographic Parity, which can be demonstrated through the example of a credit scoring model. Take, for example, a dataset of loan applicants, divided into qualified applicants (if they did actually repay the loan) and unqualified applicants (if they eventually defaulted on the loan). If African Americans have a higher rate of actual loan defaults than whites, enforcing demographic parity would result in a situation where unqualified individuals belonging to a particular demographic group of the protected characteristic with lower rates of loan repayment being assigned a positive outcome by the credit scoring model as a form of affirmative action, in order to match the percentages of those assigned a positive outcome with other demographic groups of the protected characteristic. Thus, Demographic Parity has been empirically shown to often substantially cripple the utility of the model used due to the decrease in accuracy, especially where the subject of prediction is highly correlated with the protected characteristic.

2. Equality of odds

To address Demographic Parity's problems, an alternative metric termed Equality of Odds was proposed. This metric computes both the difference between the false positive rates[62], and the difference between the true positive rates[63], of the model's decisions on the two demographic groups across the protective characteristic. [64] For instance, enforcing this metric on a model in our previous example would ensure that the rate of qualified African Americans getting a loan is equal to that of qualified whites, while also ensuring that the rate of unqualified African Americans getting a loan is equal to that of unqualified whites.

A study examining the effectiveness of Equality of Odds on the controversial COMPAS[65] algorithm which predicts recidivism of criminals, showed that although the accuracy of the algorithm was similar for both African Americans and whites, the algorithm was far from satisfying the Equality of Odds metric because the false positive rate of the algorithm's decisions

[60] M. Hardt et al., Equality of Opportunity in Supervised Learning in NIPS, 2016, 3323-3331.
[61] C. Dwork, et al., "Fairness through awareness" Proc. ACM ITCS, 2012, 214–226.
[62] A fraction of negative cases was incorrectly predicted to be in the positive class out of all actual negative cases: see S. Verma and J. Rubin, "Fairness Definitions Explained" 2018 ACM/IEEE International Workshop on Software Fairness.
[63] A fraction of positive cases was correctly predicted to be in the positive class out of all actual positive cases: see Verma and Rubin, ibid.
[64] Hardt, "Equality of Opportunity in Supervised Learning".
[65] Correctional Offender Management Profiling for Alternative Sanctions. It assesses the risk of recidivism for offenders.

was twice that for African Americans than for whites.[66] This is because in cases where the algorithm fails, it fails differently for African Americans and Whites. While African Americans are twice as likely to be predicted by the algorithm to reoffend but not actually re-offend, it was much more likely for the whites to be predicted by the algorithm to not reoffend but go on to commit crimes.

3. Equality of opportunity

Another variation is Equality of Opportunity, a weaker fairness criterion than Equality of Odds because it only matches the true positive rates across the demographic groups, without matching the false positive rate.[67] In the above example of the credit scoring algorithm, enforcing this metric would ensure qualified individuals have an equal opportunity of getting the loan, without enforcing any constraints on the model for individuals who ultimately defaulted. In some cases, Equality of Opportunity can allow the trained model[68] to achieve a higher accuracy rate due to the lack of the additional constraint.

However, it has also been found that enforcing equality only on the true positive rate will increase disparity between the demographic groups on the false positive rate.[69] In the COMPAS example above, we see a trade-off which will often be faced in machine learning classification models. Ensuring the algorithm succeeds at an equal rate in predicting African Americans and Whites as reoffending when they do actually go on to reoffend (true positive rate), results in an unequal rate of the algorithm wrongly predicting African Americans and Whites — who do not go on to reoffend — as reoffending (false positive rate). To enforce the algorithm to err at an equal rate between Whites and Africans Americans who do not actually reoffend, would almost always result in a drop in the overall accuracy of the model. This is because in the naturally occurring data, the actual rate of reoffending differs between White and African Americans.

4. Equalised calibration

Another important fairness metric to consider is equalised calibration between demographic groups. In classification models, it is often useful for a model to provide not only its prediction, but also the confidence level of its predictions. Calibration can be understood as the extent to which this confidence level provided matches with reality. Having a perfectly calibrated model would mean that if a confidence level of 0.8 is assigned to a prediction, then eight out of ten times, the predictions of the model which were assigned the confidence level of 0.8 would belong to the class predicted by the model. In recidivism models like COMPAS, risk scores are often provided along with the classification prediction of whether or not a convict will reoffend. In classification models predicting whether a borrower will default on the loan, risks scores are also provided by the model

---

[66] J. Larson et al., "How We Analyzed the COMPAS Recidivism Algorithm" *ProPublica* (23 May 2016), available at https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (last accessed 13 November 2021); also cited in D Pessach and E Shmueli, Algorithmic Fairness, arXiv:2001.09784 [cs.CY], 2020.
[67] Hardt, "Equality of Opportunity in Supervised Learning".
[68] Ibid.
[69] Pleiss, "On Fairness and Calibration".

together with the confidence level of its predictions. Where there is no perfect calibration, it is thus important that there is equalised calibration of these confidence scores between demographic groups. Otherwise, a user of the model would, for example, need to interpret a risk score for a black individual differently from a risk score for a white individual. However, as will be shown below, there is a trade-off between Equalised Calibration and Equality of Odds.

B. Trade-offs

The technical literature on fairness in machine learning has shown that there are trade-offs between the notions of fairness on both levels, namely, trade-offs between the fairness metrics for classification models (i.e. between Equalised Calibration and Equality of Odds), and trade-offs between fairness metrics and disparities in model accuracy.

1. An example of trade-offs between two fairness metrics (i.e. between equalised calibration and Equality of Odds) – *Chouldechova's Impossibility Theorem*

According to Chouldechova's Impossibility Theorem, if the prevalence (base) rates of the positive class in the demographic groups differ, it is impossible for a binary classification model to achieve all three of equalised calibration, equal false positive rates and equal false negative rates between demographic groups.[70] If a classifier has equal false negative rates between both groups, it can be mathematically derived that it will also have equal true positive rates between both groups – therefore, the Chouldechova Impossibility Theorem can be generalised to mean that a model cannot satisfy both the Equality of Odds (equal false positive rates and equal true positive rates between demographic groups) and Equalised Calibration metrics at the same time.

To put this in the context of a classification model for provision of loans, if people of colour and white individuals in the dataset do have different rates of actually defaulting on loans (the prevalence rate), it is not possible to perfectly-calibrate the credit risk scores provided by the model (such that for example, 80% of people assigned a 0.8 risk score actually default), while also having (i) the rate at which individuals predicted to default not actually defaulting (the false positive rate) to be equal between both demographic groups, and (ii) the rate at which individuals predicted to not default but actually defaulting (the false negative rate) to be equal between both groups.

Further, it was found that on the specific recidivism dataset which COMPAS was used on, enforcing an algorithm to achieve calibration would mean that there will be disparities in both the false positive and false negative rates[71] across demographic groups. On the other hand, mis-calibrated risk scores will cause discrimination to one of the demographic groups, since a user of the model would need to interpret the risks scores differently depending on the demographic group

[70] A. Chouldechova "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments" 2016 (5) Big Data 153; J. Kleinberg et. al "Inherent Trade-Offs in the Fair Determination of Risk Scores" (2017) 67 Innovations in Theoretical Computer Science 1.
[71] Pleiss, "On Fairness and Calibration".

the subject belongs to. To achieve fairness, the best course of action in such a situation is to make changes to the dataset by either collecting more data, or hopefully including more salient features in the dataset.[72]

There may be situations in which the dire consequences of false positives may differ greatly from the consequences of false negatives. In such situations, the company might choose to satisfy calibration along with only one of equalised false positive rate or equalised false negative rate. corresponding to the condition for which consequences are more dire.[73] An example to consider could be an early detection system for a chronic disease like diabetes which can be treated if detected at the early onset stage, but bearing significant long-term financial and well-being costs for the patient if left untreated till it develops into the later stage. In such a situation, the consequence of a false negative (allowing the disease to develop into the untreatable stage with long-term financial and lifestyle costs) is significantly greater than the consequence of a false positive (cost of repeated testing, or of lifestyle changes like exercise and healthy eating, aimed at reversing prediabetes), especially for lower-income minority groups. A company developing such a system might give a well-reasoned explanation for choosing to enforce calibration and an equalised low false negative rate, while forgoing equalised false positive rate.

Another example would be an experiment on an income prediction model, for deciding whether a person's income should be above $50,000. Ensuring calibration along with an equalised low false negative rate across genders would result in some employees being overpaid. This is because a false negative in such a scenario means that there are borderline cases where a male and female will each be paid less than $50,000, when in reality, one of them should have been paid more than $50,000. The company should enforce an equalised low false negative rate in a manner which would have the algorithm recommend that the company pay both of them more than $50,000[74], even if one of them is undeserving of it. For a company, this might be more tolerable than if the equalised false positive rate was chosen instead, which might result in reputational risk with some employees of a particular gender being underpaid more often than employees of another gender.

2. Trade-off between Equality of Odds and equalised accuracy

With Equality of Odds being one of the most popular and advanced metrics of fairness, it is interesting to note that there is evidence of a trade-off between Equality of Odds and equalised accuracy between the demographic groups in a dataset.[75] This was found in the dataset for the COMPAS recidivism prediction tool. In other words, this means that having the tool achieve similar levels of accuracy for African Americans and whites will result in greater differences in the false positive rate as well as the false negative rate of the tool between African Americans and whites.

---

[72] Ibid.
[73] Kleinberg, "Inherent Trade-Offs in the Fair Determination of Risk Scores".
[74] Pleiss, "On Fairness and Calibration".
[75] R. Berk et. al., "Fairness in Criminal Justice Risk Assessments: The State of the Art" (2021) 50 Sociological Methods and Research 3.

V. A FRAMEWORK FOR AI FAIRNESS REPORTING

In light of the two types of unfairness in machine learning, as discussed in Part II above – bias in classification decisions by supervised learning model, and disparities in the performance of machine learning applications across demographic groups, it is suggested that a framework for AI Fairness Reporting should consist of the following requirements: (a) disclosure of the machine learning models used; (b) disclosure of the fairness metrics and the trade-offs involved; (c) disclosure of any de-biasing methods adopted; and (d) release of datasets for public inspection or for third-party audit.

A. Disclosing All Uses of Machine Learning Models Involved

We distinguish between machine learning systems which make predictions or decisions directly affecting individuals, and machine learning systems which do not. We propose that companies should be made to furnish detailed AI fairness reports for supervised learning systems which make decisions or predictions directly and indirectly affecting individuals.

Even though our proposal does not require detailed fairness reporting for machine learning models which do not make decisions directly affecting individuals, use of any machine learning models might still bring about fairness concerns for a variety of reasons including unfair sampling. For example, crowdsourcing of data on potholes in Boston through a smartphone app which uploaded sensor data from the smartphone to the city's database resulted in more potholes detected in wealthier neighbourhoods than lower-income neighbourhoods and neighbourhoods with predominantly elderly populations, in line with patterns of smartphone usage.[76] This could have directed the use of resources on fixing potholes towards those wealthier neighbourhoods, away from poorer neighbourhoods.

A company's disclosure of all its uses of its machine learning models would allow for potential indirect implications on fairness to be flagged. Thus, companies ought to disclose all uses of machine learning models as a matter of best practice.

B. Reporting on Fairness Metrics Used and Trade-offs

Companies ought to disclose the main AI fairness metric or metrics adopted for a classification algorithm, and the reasons for its adoption. Any deliberations as to why other fairness metrics were not adopted, and how the trade-offs were navigated, also need to be explained. In light of the Chouldechova Impossibility Theorem and the trade-offs in the adoption of AI fairness metrics which have been pointed out above, along with many more which are likely to be found as research in AI fairness matures, it is important to ensure companies disclose their decisions on such trade-offs and the reasons behind it.

[76] S. Barocas et al, *Fairness and Machine Learning* (Fairmlbook.org 2019).

One way to implement and enforce explanations of deliberate omissions in reporting of AI fairness metrics is to have a robust whistleblowing policy with sufficient incentives like monetary rewards[77] as well as sanctions for companies found guilty of not explaining deliberate omissions in reporting. Employees of technology companies have not been shy to come out with concerns over the environmental and social impacts of the companies they work for. When Google allegedly forced out the co-lead of its ethical AI team over a paper which pointed out the risks of large language models which are used in recent significant enhancements to Google's core search product[78], more than 1400 Google staff members signed a letter in protest. The risks pointed out in the paper included the significant environmental costs from the large computer processing power needed to train such models, and the racist, sexist and abusive language which ends up in the training data from such models which are crawled from across the internet. Having a whistleblowing policy, coupled with an option for anonymity, would provide an accessible and effective channel for technology employees to bring omissions in reporting over such matters to light without suffering personal repercussions.

To address disparities in model performance, requiring companies to report accuracy rates (and other appropriate measures of model performance) of supervised learning models by demographic groups, instead of merely reporting an overall accuracy rate, would be a good start. However, the metric of choice for measuring model performance might not be able to capture all fairness issues, especially in machine learning applications like machine translation where the test dataset might be biased as well.

As a best practice to be encouraged, companies can consider opening up a limited interface for non-commercial use of its AI applications, where public users can probe the model to check for fairness. For example, registered users can each be allowed to upload a limited number of passages to test a translation model, or a limited number of personal selfies to test a facial recognition system.

C. Reporting on Debiasing Methods Used

Of the various approaches available for companies to satisfy the fairness metrics they have chosen for a machine learning application, each choice of approach would have different implications on trade-offs with other metrics of fairness, as well as overall accuracy, as we see below. Thus, we argue that along with the choice of fairness metrics, companies should report any interventions made to achieve fairness goals.

Methods for debiasing machine learning models have occasionally been proven to merely cover up biases with respect to a fairness metric, but not remove them. For example, methods for removing gender biases in word embeddings which reported substantial reductions in bias were

[77] For example, the US Securities and Exchange Commission will pay a monetary award to whistle-blowers who voluntarily give the SEC original information about a violation of US securities laws that leads to a successful enforcement action in which US$1 million sanctions is ordered: see US SEC, Office of the Whistleblower, Form WB-APP.

[78] K. Hao, "We Read the Paper that Forced Timnit Gebru out of Google. Here's What it Says" Technology Review (4 December 2020).

shown to have an actual effect of mostly hiding the bias, not removing it.[79] The gender bias information can still be found in the vector space distances between "gender-neutralized" words in the debiased vector embeddings, and still recoverable from them.[80] This is why the techniques for debiasing have to be reported in conjunction with the fairness metrics, to prevent such "over-optimising" on the chosen fairness metric without serving the actual goal of fairness. It is important to note that the debiasing techniques used can be reported with little to no revelation about the AI model itself. Thus, companies should have no excuse for not reporting on the basis of protecting their trade secrets.

1. Pre-processing methods

Pre-processing methods make changes to the dataset before the machine learning algorithm is applied. As discussed earlier, prevalence rates of the target variable of prediction, say the occurrence rate of recidivism, may differ across demographic groups. Methods of rebalancing the dataset could involve relabelling some of the data points (an example of which is changing the label of a random sample of men who failed on parole to a success), or assigning weights to the data points and weighing less represented groups in the dataset more heavily. As intuition would readily tell us, rebalancing would likely lead to a loss in accuracy. There are other more sophisticated methods of pre-processing, which can be optimised in a manner which changes the values of all predictive features in the dataset while still preserving as much 'information as possible[81], but it remains to be seen whether such methods will result in other trade-offs.

Deep learning models learn rich representations of the data they are fed. Thus, the need to learn fair representations emerged, with the underlying idea being that if representations of instances from different demographic groups are similar to each other, then predictive models built on top of them will make decisions independent of group membership.[82] A method of learning such fair representations in practice is to build another separate model, called an adversarial discriminator, which is tasked with predicting the group membership of data instances[83]. The deep neural network tasked with creating the data representations would then create "fair representations" in a manner which makes it hard for the adversarial discriminator model to predict the group membership of the data from the representation. This competition between the two models pitted against each other during the training process results in data representations which carry little information about group membership of the individual. However, it has been proven both theoretically and empirically that machine learning fair representations present a trade-off between Demographic

[79] H. Gonen and Y. Goldberg, "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them" NAAACL 2019.

[80] The study found that the way the gender bias of word embeddings was defined is merely a way of measuring the gendered-nature of the word embedding, and not determinative of gender bias in the word embedding. Thus, even though methods of debiasing tried to cure the word embedding on that measure of bias, other experiments revealed that the word could still be associated with embeddings of other words biased towards the particular gender.

[81] J.E. Johndrow and K. Lum, "Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction" (2017) arXiv:1703.04957 [stat.AP].

[82] H. Zhao and G.J. Gordon, "Inherent Tradeoffs in Learning Fair Representations" 2015 NIPS.

[83] T. Adel et. al., "One-network Adversarial Fairness" 2019 AAAI.; A. Beutel et. al., "Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. 2017 arXiv preprint arXiv:1707.00075; H. Edwards and A. Storkey. "Censoring Representations with an Adversary" 2015 arXiv preprint arXiv:1511.05897; C. Louizos, et. al., "The Variational Fair Autoencoder" 2015 arXiv preprint arXiv:1511.00830.

Parity and accuracy of the models trained, when the rate of occurrence of the positive class in the dataset differs between groups.[84]

2. *In-processing*

In-processing makes fairness adjustments during the process of the machine learning model making predictions. This could involve changes to the model so that a specified fairness goal is taken into account.[85]

3. *Post-processing*

Post-processing involves changing the predictions of a machine learning model to achieve better results on a fairness metric. This can be done through randomly reassigning the prediction labels of the model.[86] However, the result of such reassignments could be that the overall classification accuracy of the model is brought down to match that of the demographic group for which accuracy was the worst. Besides, having individuals being randomly chosen to be assigned a different outcome might raise individual fairness concerns when similar individuals are treated differently. There might also be ethical considerations when such methods are used in sensitive domains like healthcare.

Technical research on the implications of debiasing techniques is still nascent, though there is evidence of consequences for both model accuracy and trade-offs with other fairness goals not taken into account in the technique applied. Making it mandatory for companies to transparently report any debiasing interventions made would allow public scrutiny and academic research to flag potential implications, intended or unintended, of the procedure chosen.

D. Release of Datasets for Public Inspection or for Third-Party Audit

Ideally, companies should release all datasets used for the training of machine learning models to the public.[87] However, it is understandable that significant investment is often required on the part of companies to collect and curate such datasets in order to obtain a competitive advantage, so companies might be reluctant to share such data. Also, some datasets might also contain trade secrets, confidential information, and the private data of users. It might not always be feasible to

[84] Zhao and Gordon, "Inherent Tradeoffs in Learning Fair Representations".
[85] T. Kamishima et. al., "Fairness-aware Learning Through a Regularization Approach" 2011 11th IEEE International Conference on Data Mining Workshops.
[86] Hardt, "Equality of Opportunity in Supervised Learning". For example, this means that the labels of a random subset of data instances which the classification model assigned to the positive class would be reassigned to the negative class.
[87] Instead of setting an arbitrary threshold, the degree of such data release is best left to be decided by the company on a case-by-case basis through a comply-or-explain approach, taking into account considerations like user privacy and trade secrets protection.

completely prevent data re-identification from the release of anonymised data. Thus, the release of datasets should not be mandated, but best left to a comply or explain basis.[88]

However, in cases where the dataset is not released, we propose that a requirement be set for an independent third-party audit to be done on the dataset. This audit can flag any potential problems of bias in data labelling from operational practice, or underrepresentation of specific demographic groups. The audit report should be made public together with the AI Fairness Report in the company's public disclosures.

Much can be done to encourage companies to release their datasets, and the availability of such data would aid the progress of research into AI fairness. First, for companies to preserve their competitive advantage, the release of such datasets does not need to be made under an open-source license[89]. A new standard data release license, similar to the non-commercial and no derivatives licenses for research data[90], can be created such that the use of the data is limited to inspection for fairness concerns. Admittedly, enforcement of such a license can be a problem if it is possible for models to be trained on the released data with little risk of detection by the data owner.

Second, companies might be concerned about the impact on user privacy should such datasets contain user information, and potential liability from breaches of data protection regulations. Data protection authorities can consider providing a safe harbour for datasets released to facilitate AI fairness, as long as anonymisation procedures under guidelines issued by data protection authorities are followed to reduce the risk of data re-identification.

One major limitation to note on the release of anonymised datasets is how much it correctly represents the nature of the original dataset, especially if modifications[91] to values in the dataset had to be made to prevent re-identification of individuals. It might be possible that the anonymised dataset released might in turn be a misrepresentation of fairness in the original dataset. It might be helpful to mandate that any data anonymisation procedures applied to the released data be declared by the company to mitigate this concern.

Apart from releasing the proprietary data used for model training, the company should also disclose any use of open-source datasets and pre-trained models from third parties. This would allow the public to consider whether any known biases in such open-sourced datasets and pre-trained models might be carried into the company's AI models.

---

[88] MacNeil and Esser, "The Emergence of 'Comply or Explain' as a Global Model for Corporate Governance Codes".
[89] Project Open Data, https://project-open-data.cio.gov/open-licenses/ (last accessed 13 November 2021).
[90] OpenAire, https://www.openaire.eu/research-data-how-to-license/ (last accessed 13 November 2021).
[91] Perturbations to datasets through the addition of random noise is one way of reducing the risk of re-identification in anonymised datasets released for research purposes, but it is unclear whether such methods should be used on a dataset released for fairness reporting. See R.V. Atreya et. al., "Reducing Patient Re-identification Risk for Laboratory Results Within Research Datasets" (2013) 20 J. Am. Med. Inform. Assoc. 9.

IV. APPLICATION OF AI FAIRNESS REPORTING FRAMEWORK TO TWO CASE STUDIES

A. Goldman Sachs' Credit Profiling Model on Issuance of Apple Card

We consider the case of Goldman Sachs' credit profiling of applicants for the Apple Card. A technology entrepreneur, David Heinemeier Hansson, called out Goldman Sachs' Apple Card program for gender-based discrimination through the use of what he termed a "black-box algorithm"[92]. Claiming that although he and his wife filed joint tax returns and live in a community-property state, he received a credit limit that was 20 times higher than what his wife was offered. Concerns were also raised that "the Bank relied on algorithms and machine learning for credit decision-making and complained that an Apple Card customer service agent could not explain these algorithms or the basis for the difference in credit limits".[93] Apple's cofounder Steve Wozniak also claimed that he had 10 times the credit limit of his wife on the Apple Card, even though they shared all assets and accounts.

We now turn to look at how AI Fairness Reporting under our framework could be retroactively applied in this case. Even though no fair lending violations were found by the New York State Department of Financial Services, we argue that had this reporting been done, the transparency and communication issues flagged[94] by the New York State Department of Financial Services report could have been at least mitigated, if not avoided entirely.

1. Disclosing all uses of machine learning models

Under the proposed AI Fairness Reporting framework, Goldman Sachs ought to have disclosed all its uses of machine learning models as a matter of best practice. Even the use of machine learning models which do not make directions or predictions directly affecting individuals would need to be declared – this would include internal risk management models which predict the health of Goldman Sachs' lending business. If the internal risk models consistently predict a high-risk exposure to Goldman Sachs' lending business ahead of a holiday specific to one demographic group, causing Goldman Sachs' to generally tighten credit lending ahead annually at this time of the year in line with an increase in credit needs from this demographic group, this could raise fairness considerations.

The machine learning models used in Goldman Sachs relating to the Apple Card Program, which directly affect individuals, include more than just the credit scoring model. Machine learning models deployed on Goldman Sachs' consumer-facing platforms, which determine whether to advertise or recommend the Apple Card to a particular user, need to go through detailed fairness reporting as well.

---

[92] S. Perez, "New York's Department of Financial Services says Apple Card Program Didn't Violate Fair Lending Laws" *Techcrunch* (24 March 2021).
[93] New York State Department of Financial Services, Report on Apple Card Investigation (March 2021), 4.
[94] Ibid.

2. Reporting on fairness metrics used

The choice of fairness metrics needs to take into account the social and legal contexts of the machine learning application. For credit lending decisions, the Equal Credit Opportunity Act and state laws in the US apply to Goldman Sachs' Apple Card programme. The sex of credit applicants cannot be taken into account in the credit decisions, and two categories of discrimination are recognised: disparate treatment and disparate impact. Debiasing a machine learning model together with the disclosure of group fairness metrics would reveal that protected characteristics like sex had been taken into account. This could contravene the disparate treatment requirement since it disallows the intended use of protected characteristics.

At the same time, to examine disparate impact, the Consumer Examinations Unit of the New York State Department of Financial Services applied regression analysis on the Bank's Apple Card underwriting data for nearly 400,000 New York applicants, covering applications dating from the launch of Apple Card until the time of the initial discrimination complaints. It did not state if any specific fairness metric was used, but the regression analysis would have measured the independence between sex and the credit decisions made.[95] The Department found that the Bank had a fair lending program in place for ensuring its lending policy "did not consider prohibited characteristics of applicants and would not produce disparate impacts", with an "underlying statistical model".[96] The New York State Department of Financial Services, in its investigation report[97], also found that "women and men with equivalent credit characteristics had similar Apple Card application outcomes". This seems to allude to a notion of individual fairness also being applied in the report.

In such a situation, Goldman Sachs would have to choose both a group fairness metric and an individual fairness metric to report on.[98] It is highly likely that there would have been trade-offs between the chosen group fairness metric and the individual fairness metric. In the context of this case, enforcing the algorithm to give a high credit rating at an equal rate to men and women who do not ultimately default on payments might result in individuals with highly similar profiles being given a different credit rating. This can happen when for example, men have more borderline cases than women, and in order to equalise the rate at which a high credit rating is predicted between men and women who did not ultimately default, highly similar borderline profiles of men might be assigned different outcomes. All metrics used in arriving at the operational model should thus be reported to transparently show how these trade-offs were navigated in the final model used.

3. Reporting on debiasing methods used

[95] Another related fairness metric, also termed disparate impact, is based on the fourth-fifths rule advocated by the US Equal Employment Opportunity Commission. However, it is unclear whether this metric is apt to be applied in a credit lending situation.

[96] Although the credit decisions were found by the Department not to violate the law, the news scandal and associated reputational fallout could have been avoided had there been greater transparency upfront based on the fairness framework.

[97] New York State Department of Financial Services, Report on Apple Card Investigation (March 2021), 4.

[98] Fleisher, "What's Fair About Individual Fairness?".

What is left completely missing in both the investigation report and subsequent public relations efforts by Goldman Sachs on the Apple Card program is the specific debiasing methods used to arrive at the fairness outcomes.

Existing laws like the Equal Credit Opportunity Act serve to protect consumers from discriminatory lending on protected characteristics, so the investigation report's finding that no fair lending laws have been violated serves little to inform other stakeholders on how the use of the machine learning model affects them. Stakeholders of Goldman Sachs would be interested to know how much the debiasing methods used (if any) would have an impact on accuracy of the credit scoring model as this will affect the business and operations of Goldman Sachs, which will in turn impact its financial performance and reputation. Researchers can further concentrate their study of the implications of such debiasing techniques used in practice, in the specific context of credit scoring, given that the full implications of debiasing techniques are still an under-researched area. Credit applicants themselves would want to know how such debiasing techniques might potentially affect them, beyond a report of mere general compliance with the law.

4. Release of datasets for inspection

We make reference to the German Credit Dataset, a standard credit scoring dataset used in machine learning fairness research, as an indication that it might have been possible for Goldman Sachs to have released an anonymised dataset of applicants to its Apple Card program. The German Credit Dataset consists of 1000 individuals drawn from a German bank in 1994. Protected characteristics in the dataset include gender and age, along with 18 other attributes including employment, housing and savings.

A third-party audit of datasets used to train any machine learning models used for credit scoring in the Apple Card program would have been required, if there was no release of a public dataset. These datasets could include Goldman Sachs' historical data on setting credit limits on other similar credit programs, and any bias in those datasets could have carried over to the Apple Card program if models were trained on the data.

However, even if Goldman Sachs deemed that the release of such a dataset would pose significant risks for client privacy, it could have been more transparent by giving a comprehensive listing of the attributes which are taken into account in its credit scoring model. That would have reduced misunderstandings as to why seemingly similar individuals were offered different credit limits. Explanations given[99] in the Department's report as to why spouses with shared bank accounts and assets in this incident were given different credit outcomes included obscure attributes which might not have been considered by a layman. These included "one spouse was named on a residential mortgage, while the other spouse was not", and "some individuals carried multiple credit cards and a line of credit, while the other spouse held only a single credit card in his or her name". Even if an applicant had referred to public education materials which were released by Goldman Sachs after this incident—the "A Closer Look at our Application Process" portion of the website which

[99] New York State Department of Financial Services, Report on Apple Card Investigation (March 2021), 10.

provides a snapshot of the data the Bank draws upon in setting credit terms[100]—the applicant would not know the attributes that Goldman Sachs took into account in its credit scoring model.

B. Wrongful Arrest Attributed to False Positive Match by Dataworks Plus' Facial Recognition System

We next consider the case where the facial recognition technology by a US company Dataworks Plus resulted in a wrongful arrest in the US state of Michigan. Robert Julian-Borchak Williams, an African American man, was wrongfully accused of shoplifting due to a false positive match by Dataworks Plus' facial recognition software.[101]

This culminated in the request by Senator Sherrod Brown of Ohio for Dataworks Plus to provide information to the US Congress on questions including whether the company plans to impose a moratorium on the use of its facial recognition technologies by law enforcement, the factual basis behind marketing claims by the company on the reliability and accuracy of its facial recognition system, and whether there is an executive responsible in the company for facilitating conversations on ethical decision making.[102] Keeping in mind that Dataworks Plus brands itself as a "leader in law enforcement and criminal justice technology"[103], with the facial recognition system FACE Plus being one of its key offerings, imposing such a moratorium will have a substantial impact on its financial revenue.

This case is different from the previous case, in that the creator of the facial recognition system is not the user of the deployed system – the Detroit police department was the user. Also, there is a nuanced difference here on the allegation of unfairness. This is not a problem of disparate outcomes across a protected characteristic, but of the AI system having a different level of accuracy for different demographic groups. Here, the facial recognition system is matching facial snapshots from crime scene video surveillance to a 50 million Michigan police database of driver license photographs and mug shots in order to generate potential suspect candidates. The allegation is that the quality of candidates produced by the facial recognition system is worse when it comes to people of colour.

This allegation is not unfounded, given the findings of studies preceding the incident, conducted on commercial facial recognition systems. In a Massachusetts Institute of Technology study[104] it was found that the error rate for light-skinned men is never worse than 0.8 percent, but 34.7 percent for dark-skinned women. According to the study, although researchers at a major U.S. technology company claimed an accuracy rate of more than 97 percent for a face-recognition system they had designed, the data set used to assess its performance was more than 77 percent male and more than 83 percent white. A National Institute of Standards and Technology study[105] covered 189 software algorithms from 99 developers, which make up the majority of the industry in the US. The study

---

[100] Ibid, 9.
[101] K. Hill, "Wrongfully Accused by an Algorithm" *New York Times* (3 August 2020).
[102] Ibid
[103] Dataworks Plus, http://www.dataworksplus.com/index.html (last accessed 13 November 2021).
[104] L. Hardesty, "Study finds gender and skin-type bias in commercial artificial-intelligence systems" *MIT News Office* (11 February 2018).
[105] P. Grother et. al., "Face Recognition Vendor Test" (FRVT) Part 3: Demographic Effects (NISTIR 8280, 2019).

used four collections of photographs containing 18.27 million images of 8.49 million people from operational databases provided by the State Department, the Department of Homeland Security and the FBI. It found that for one-to-many matching systems which are commonly used in suspect identification systems, there was a higher rate of false positives for African American women, although the study also caveated that not all algorithms give this high rate of false positives across demographics in one-to-many matching, and systems that are the most equitable are also amongst the most accurate. By the account of the Detroit Police Chief, Dataworks Plus' facial recognition system misidentifies 96 percent of the time.[106] From the results of the NIST study, this might indicate that the allegation that it has a higher rate of false positives for African Americans is a reasonable one to make.

Applying our AI Fairness Reporting Framework on Dataworks Plus, we argue that the process would have better enabled Dataworks Plus to identify problems with its facial recognition system and would have allowed the civilian oversight board[107] in Detroit to better evaluate the adoption of the system.

1. Disclosing all uses of machine learning models

Dataworks Plus, being a provider of software systems rather than a user, would need to disclose all the uses of machine learning models in the various software solutions it provides. There might be multiple machine learning models in a single software system. For example, a facial recognition system might have an image classification model to first classify the race of the subject of a facial image, before applying a matching algorithm built specifically for image subjects belonging to that particular race.

We do note that there might be concerns about the protection of trade secrets, with mandating the disclosure of machine learning model use. However, there can be a degree of flexibility afforded to the company with regards to the granularity of disclosure – the disclosure can range from the general class of machine learning model, to the specific model used. It will be hard for a company to justify why such a requirement, with flexibility afforded, cannot be imposed on companies – especially when balanced against the interests of stakeholders like shareholders, potential customers and individuals whose lives might be affected by the use of the models.

2. Reporting on fairness metrics used

The NIST Face Recognition Vendor Test report[108] studied the differences in false positives and false negatives between demographic groups in the dataset, along the lines of sex, age and racial background. We recommend that these two metrics are apt for use in AI Fairness Reporting by the company. This would be a holistic representation of how well the facial recognition system performs, in stark contrast to marketing materials on Dataworks Plus' website highlighted by

[106] J. Koebler, "Detroit Police Chief: Facial Recognition Software Misidentifies 96% of the Time" *Tech By Vice* (30 June 2020).
[107] Detriot Board of Police Commissioners, https://detroitmi.gov/government/boards/board-police-commissioners.
[108] Grother, "Face Recognition Vendor Test".

Senator Brown, which vaguely described the facial candidates produced by the Face Plus software system as 'accurate and reliable'.

When the incident of wrongful arrest was first reported on the New York Times, the General Manager of Dataworks Plus, Todd Pastorini, was cited as claiming that checks which Dataworks Plus does when they integrate facial recognition systems from subcontractors are not "scientific", and that no formal measures of the systems' accuracy or bias were done. All these bad press for the company, and its associated reputational risks could have been avoided had a fairness study been conducted and reported on by the company. The Dataworks Plus facial recognition software used by the police in Michigan includes components developed by two other companies, NEC and Rank One Computing.[109] The NIST study[110] conducted the year before the incident on over a hundred facial recognition systems including those developed by these two companies, had found that African American and Asian faces were ten to a hundred times more likely to be falsely identified than Caucasian faces.[111]

However, one more nuance needs to be appreciated in this situation where the developer of the AI system is not the end user – the prediction outputs of the AI system need to be interpreted and acted upon by the user. In this case, the system provided a row of results generated by the software from each of the two companies, NEC and Rank One Computing, along with the confidence scores of each candidate match generated.[112] It is up to the investigator to interpret these matching candidates, along with the associated confidence scores, before making a judgement on whether to proceed with any arrest. The outputs of the AI system are thus defended by law enforcement and software providers like Dataworks Plus as mere investigative leads, and not conclusive on arrest decisions. In such a situation, assuming proper use of the system, the presence of false positives is not as detrimental as it might be sensationalised to be. Thus, explanations about the context of the AI system's use, and guidance on how the reported fairness metrics should be interpreted, would be helpful if included in the AI Fairness Reporting.

3. Reporting on debiasing methods used

This is a case where there is a clear risk that the use of debiasing methods could create other problems of concern. A study[113] by computer scientists at the Florida Institute of Technology and the University of Notre Dame showed that facial recognition algorithms return false matches at a higher rate for African Americans than white people, unless explicitly recalibrated for the African American population. However, such recalibration would result in an increase in false negatives for white people if the same model is used, which means it will make it easier for the actual white culprits to evade detection by the system. Using different models, however, would require a separate classification model for choosing the appropriate model to use, or require the police to

[109] Hill, "Wrongfully Accused by an Algorithm".
[110] Grother, "Face Recognition Vendor Test".
[111] Hill, "Wrongfully Accused by an Algorithm".
[112] Ibid
[113] A. Harmon, "As Cameras Track Detroit's Residents, a Debate Ensues Over Racial Bias" *New York Times* (9 July 2019); K.S. Krishnapriya et. al., "Characterizing the Variability in Face Recognition Accuracy Relative to Race" CoRR abs/1904.07325 (2019).

exercise judgment which might introduce human bias[114]. It is thus important that the methods used to address bias are disclosed in order for observers to anticipate and flag any potentially inadvertent problems the models create.

4. Release of datasets for inspection

The datasets contain the photographs of individuals, which make anonymisation without removing important information in the data practically impossible. However, the metadata of the subjects can be released, and reference can be made to the metadata information used in the NIST study[115] indicating the subject's age, sex, and either race or country of birth. This transparency with regards to metadata information would allow for underrepresentation of demographic groups in the dataset to be detected and flagged by observers, and is in our view sufficient for the purposes of disclosure.

VII. CONCLUSION

Thus far, regulators and the legal literature have been treating fairness as a principle of AI governance, but shy away from prescribing specific rules on how this principle should be adhered to. That approach may be justified in view of the technical uncertainty over how fairness in AI should work in practice, and the myriad considerations and contexts in which it operates. However, technical progress in AI fairness research has highlighted the issues arising from the fairness metrics used and the important trade-offs in the deployment of AI, including between AI fairness metrics as well as accuracy. There are also reported incidents of bias in artificial intelligence systems which have captured the public consciousness, leading to backlash against companies in the form of employee walkouts, resignations of key executives[116] and media scrutiny[117].

Reflexive regulation in the form of AI Fairness Reporting according to the framework proposed in this paper encourages companies to take the necessary steps to ensure the fairness of AI systems used or sold, while empowering stakeholders of a company with adequate information to flag potential concerns of unfairness in the company's AI systems. It also affords companies with a measure of flexibility to take into account other considerations like user privacy and protection of trade secrets when they are reporting on AI fairness.

One limitation of the AI Fairness Reporting framework is that it only captures the fairness outcomes of machine learning models at a snapshot at the time of reporting. Even if companies are subject to such reporting on an annual basis, it is at best an ex-post monitoring mechanism when shifts in the nature of the data happen between reporting periods. Companies might also

[114] Hill, "Wrongfully Accused by an Algorithm".
[115] US Department of Commerce, National Institute of Standards and Technology, NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software (19 December 2019).
[116] J. Dastin and P. Dave, "Google AI Scientist Bengio Resigns After Colleagues' Firings: Email" *Reuters* (7 April 2021).
[117] R. Mac, "Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men" *New York Times* (14 September 2021).

pushback on how the AI Fairness Reporting would create an onerous burden for companies using AI, and holds the use of AI to a higher standard of interrogation than that for human decision makers. However, it is important to note the opportunity which AI opens up for us to combat unfairness, which was not available with human decision makers. Despite the complaints about the opacity of AI, AI is still far more transparent through the methods outlined in the proposed framework than the conscious (and unconscious) thoughts in the brain of a human decision maker. As compared to our ability to inspect the datasets used to train an AI model, it is much harder to access and assess all the experiences in the lifetime of a human decision maker which might influence how a decision is made. Similarly, while explicit debiasing methods are applied to an AI model in order to achieve the reported AI fairness metrics, it is harder to assess how a human decision maker corrects, and potentially overcorrects, for the biases of which they are aware. Businesses should see the increased compliance costs as part of the bargain for accessing the benefits of AI. We can look to the progress of climate-change reporting in the UK, which has now been made mandatory[118], in the hope that efforts to ensure companies act more responsibly towards their stakeholders, such as the proposed AI Fairness Reporting, can have a similar traction.

[118] The Financial Conduct Authority has already implemented comply or explain TCFD disclosure for LSE premium-listed issuers, and intends to mandate TCFD disclosure requirements for large private companies, banks, and other companies: HM Treasury, 'Interim Report of the UK's Joint Government Regulator TCFD Taskforce' (2020), 14-16, available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/933782/FINAL_TCFD_REPORT.pdf (last accessed 13 November 2021).