Cambridge Law Journal, **81(3)**, November 2022, pp. 610–644 © The Authors, 2022. Published by Cambridge University Press on behalf of The Faculty of Law, University of Cambridge. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.
doi:10.1017/S0008197322000460

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 requires 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: (1) disclosure of all machine learning models usage; (2) disclosure of fairness metrics used and the ensuing trade-offs; (3) 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, equality, 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, as to 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, the editor, Louise Gullifer, and the two anonymous referees for their insightful comments. The usual disclaimers apply.

[1] See e.g. Organisation for Economic Co-operation and Development, "Recommendation of the Council on OECD Legal Instruments Artificial Intelligence" (2019), available at https://oecd.ai/en/ai-principles (last accessed 15 June 2022); "State of Implementation of the OECD AI Principles: Insights from

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 consist of 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 stakeholders to assess the full extent of fairness risks faced by the company in the provision and use of AI. Besides, investors with increased awareness of sustainable investing would want 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

National AI Policies" (2021) OECD Digital Economy Papers No. 311, available at https://www.oecd.org/digital/state-of-implementation-of-the-oecd-ai-principles-1cd40c44-en.htm (last accessed 15 June 2022).

[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*, available at https://www.ft.com/content/a5970b6c-e731-45a7-b75b-721e90e32e1c (last accessed 15 June 2022).

[4] For the industry benchmark, see the Global Reporting Initiative (GRI) Standards, available at https://www.globalreporting.org/how-to-use-the-gri-standards/ (last accessed 21 June 2022).

[5] United States Securities and Exchange Commission, "Form 10-K: Microsoft Corporation" (annual report for the fiscal year ended 30 June 2021), 28, available at https://www.sec.gov/Archives/edgar/data/0000789019/000156459021039151/msft-10k_20210630.htm (last accessed 15 June 2022). It is striking that one of the world's top three largest and most influential technology companies merely devotes fewer than 10 sentences to the risks of AI in its more than 100-page annual report.

[6] See e.g. I. Chiu and E. Lim, "Managing Corporations' Risk in Adopting Artificial Intelligence: A Corporate Responsibility Paradigm" (2021) 19 Washington Univ. Global Stud. 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 optimisation, 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", available at https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2020 (last accessed 15 June 2022).

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

Reporting Initiative (GRI) standards in sustainability reporting.[7] Sustainability reporting rules (and practice notes) require (or advise) companies to describe both the reasons and the process of selecting material ESG factors.[8] In a similar way, companies should be required to report on the AI fairness metrics that they have adopted for the algorithms and the reasons for adoption, in a manner which will be useful for public scrutiny and debate by stakeholders, regulators and civil society.

Unfortunately, current guidance on Data Protection Impact Assessments (DPIA) under the General Data Protection Regulation (GDPR) does 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 on a comply-or-explain basis based on our proposed reporting framework.[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 common sources of unfairness are 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 addressing of what the disclosure contents of the AI Fairness Reporting framework should consist. Fifth and finally, the proposed Reporting framework is applied to two case studies.

The structure of this article is as follows. Section II provides three reasons for having the AI Fairness Reporting framework: (1) to enable investors and stakeholders to have a better understanding of the potential legal liability risks due to contravention of applicable legislation; (2) to address investors' and stakeholders' sustainability-related expectations concerning

[7] GRI, available at https://www.globalreporting.org/how-to-use-the-gri-standards/ (last accessed 21 June 2022).

[8] See e.g. Regulation (EU) 2019/2088; European Commission, "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).

[9] A. Kasirzadeh and D. Clifford, "Fairness and Data Protection Impact Assessments" (2021) AEIS '21 146; M.E. Kaminski et al., "Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations" (2021) 11 I.D.P.L. 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), 154–57.

[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 e.g. 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.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

the company's business and operations; and (3) to address inadequacies in the DPIA under the GDPR.

Section 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.[11] The unfairness can also arise from disparities in the performance of machine learning systems with respect to data related to different demographic groups.

Section 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.

Section V advances a framework for AI Fairness Reporting, the proposed reporting obligations of which should include: (1) disclosure of all uses of machine learning models; (2) disclosure of the fairness metrics used and the ensuing trade-offs; (3) disclosure of the de-biasing methods used; and (4) release of datasets for public inspection or for third-party audit.

Section 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,

[11] See Section III(A)(2).

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

pregnancy and maternity, race, religion or belief, sex and sexual orientation.[12]

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 creditworthiness 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.[13]

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".[14] 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[15] is prohibited under Article 9(1) of the GDPR, unless one of the exceptions in Paragraph 2 is satisfied. This

[12] See Equality Act 2010, s. 4.

[13] 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 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.

[14] 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 Regulation (EU) 2016/679 (OJ 2016 L 119 p. 1) (GDPR).

[15] 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."

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

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[16] unless the engineer fulfils one of the Paragraph 2 exceptions such as consent. However, it has been argued that genuinely free consent cannot be obtained in this case, because a refusal to grant consent could result in the individual suffering a higher risk of discrimination, such as being denied the opportunity to apply for a job.[17]

Even if special category data are not processed directly, other data categories in the dataset might be used as proxy information to infer the special category data. The law is unclear as to when the existence of multiple proxy information available in the dataset, which allow for special category data to be inferred, would be deemed by the regulator to amount to special category data. The UK's Information Commissioner's Office guidelines on special category data state that the question of whether proxy information, which allows special category data to be inferred, will be deemed by the regulator as special category data depends on the certainty of the inference, and whether the inference was deliberately drawn.[18] Courts, in interpreting this provision, are likely to distinguish between (1) an explicit inference of special category data made by an algorithm in its final prediction and (2) algorithms which make predictions correlated with special categories without actually making the inference that the person in question possesses the special characteristics.[19] In addition to the latter case, we think algorithms which are provided with data correlated with special categories would belong to that category too, and this latter case should not trigger Article 9.

3. *Domain-specific Legislation in the US*

The US has domain-specific legislation in a variety of areas where machine learning is now applied, for example, the Fair Housing Act[20] and the Equal Credit Opportunity Act,[21] which list protected characteristics which are

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

[17] 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", [13], available at https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf (last accessed 15 June 2022) which states that consent will not be valid if the data subject will endure negative consequences if they do not consent.

[18] 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).

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

[20] Sec. 804 42 U.S.C. 3604.

[21] 15 U.S.C. 1691.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

similar to those listed in the UK Equality Act. 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.[22] Disparate treatment comprises either formal disparate treatment of similarly situated people or treatment carried out with the intent to discriminate. Disparate impact refers to practices that are superficially neutral but have a disproportionately adverse impact on groups with protected characteristics. Disparate impact is not concerned with intent, but to establish it, three questions need to be asked. First, whether there is a disparate impact on members of a group with a protected characteristic; second, whether there is a business justification for that impact; and finally, whether there are less discriminatory ways of achieving the same result.[23] The US Equal Employment Opportunity Commission advocates for a four-fifths rule,[24] namely, that 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.

Our proposed AI Fairness Reporting would allow investors, stakeholders and regulators to better assess whether sufficient work has been done by the company to comply with such regulations. Reporting on the fairness of AI models would also help to inform investors and 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 a loss of talent.

B. Sustainable Investments

There has been a rapid growth in sustainable investments in the last few years. This has resulted in the incorporation of various ESG-related concerns or objectives into investment decisions. Globally, assets under management in ESG mutual funds and exchange-traded funds have grown from $453 billion in 2013 to $760 billion in 2018 and are expected to continue growing.[25] It is plausible that AI fairness considerations are already being taken into account by such ESG funds, (or will be in the near future) as part of their compliance with ESG reporting requirements. There is already work being done by investment funds on establishing a set of requirements including non-bias and transparency of AI use.[26] This

[22] S. Barocas and A.D. Selbst, "Big Data's Disparate Impact" (2016) 104 C.L.R. 671.
[23] Ibid.
[24] The U.S. EEOC. Uniform Guidelines on Employee Selection Procedures, 29 CFR Part 1607.
[25] BlackRock, "Sustainability: The Future of Investing", available at https://www.blackrock.com/us/individual/literature/whitepaper/bii-sustainability-future-investing-jan-2019.pdf (last accessed 15 June 2022).
[26] Hermes Investment Management, "Investors' Expectations on Responsible Artificial Intelligence and Data Governance", available at https://www.hermes-investment.com/wp-content/uploads/2019/04/investors%E2%80%99-expectations-on-responsible-artificial-intelligence-and-data-governance.pdf (last accessed 15 June 2022).

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

set of requirements could then be used by investment funds to evaluate the use of AI by a company.

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.[27] Stakeholder capitalism is premised on the idea that the stock market misvalues intangibles that affect stakeholders, such as employee satisfaction.[28] Therefore, it emphasises that corporate directors and executives should make decisions in a manner which takes into account the interests of stakeholders other than shareholders, such as customers, employees and society at large. A natural extension of the considerations that corporate directors are required to take into account in order to make decisions which accord with stakeholder capitalism would be whether AI products and services used or sold by the company are fair towards potential job applicants, employees, customers and 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".[29] This reference to the "rights and freedoms of natural persons" is to be interpreted as 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.[30] 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 in Recital 71 which specifically highlights processing operations as including those of a "profiling" nature such as "analysing or predicting aspects concerning performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements". Further, Article 35(3)(b) relates to "processing on a large scale of special categories of data referred to in

[27] K. Schwab and P. Vanham, *Stakeholder Capitalism: A Global Economy that Works for Progress, People and Planet* (Hoboken 2021); McKinsey, "The Case for Stakeholder Capitalism", available at https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-case-for-stakeholder-capitalism (last accessed 15 June 2022).

[28] Ibid.

[29] GDPR, Article 35.

[30] 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.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

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".

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 Recitals 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".[31] The methodology of the DPIA is left up to the data controller. Even though guideline criteria are provided,[32] they make no reference to any fairness metrics and de-biasing techniques[33] which have emerged in the technical machine learning literature.[34]

Although previous work on biased hiring algorithms called for DPIA reports to be made available publicly,[35] there is no current requirement under the GDPR for such DPIA reports to be made public. Moreover, we do not think DPIA reports in their current form as defined under the GDPR and their guidance documents adequately serve the needs of AI Fairness Reporting because the DPIAs do not require the disclosure of fairness metrics and the de-biasing methods used.[36]

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[37] of machine learning models: supervised learning, unsupervised learning and reinforcement learning. Supervised learning models are trained on data examples labelled with the decision which needs to be made. These labels are created either by manual human labelling or

[31] Ibid.

[32] Ibid, "Annex 2 – Criteria for an acceptable DPIA".

[33] De-biasing techniques adjust a machine learning model in a manner such that the results of the model will satisfy a fairness definition adopted by the engineer. We discuss some of these techniques in greater detail in Section V(C).

[34] Kasirzadeh and Clifford, "Fairness and Data Protection Impact Assessments".

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

[36] See Section V for an analysis of the de-biasing methods.

[37] K.P. Murphy, *Machine Learning: A Probabilistic Perspective* (Cambridge and London 2012).

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

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. Reinforcement learning models use reward or punishment signals to learn how to act or behave. These models are distinct from supervised and unsupervised learning models. In our discussion, we focus primarily on supervised learning models. These have, so far, brought about the most legal and policy concerns surrounding fairness.

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 dataset.[38] Supervised learning models are dependent on the labels given to data in the training set. If the organisation has been making unfair decisions reflected in the training dataset, such unfairness will be included 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.[39] Automatic essay grading models might then be trained on a dataset of essays with the corresponding scores assigned by such human essay graders, thus incorporating 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 their 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[40] 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 numerical representations known as vector embeddings, which can capture meaning and semantic relationships

[38] T. Kamishima et al., "Fairness-aware Classifier with Prejudice Remover Regularizer" in P. Flach, T. Bie and N. Cristianini (eds.), *Machine Learning and Knowledge Discovery in Databases* (Berlin and Heidelberg 2012), 35.

[39] S. Barocas, M. Hardt and A. Narayanan, *Fairness and Machine Learning* (2019), available at fairmlbook.org (last accessed 15 June 2022). See also R.N. Hanna and L.L. Linden, "Discrimination in Grading" (2012) 4 Amn. Econ. J. 146.

[40] American Standard Code for Information Interchange.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

between words through their distance and directional relationship with vector embeddings 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 a technique known as representation learning to learn the representations as vector embeddings instead. In the context of natural language processing, representation learning is done by training on large datasets like Common Crawl,[41] using the frequency of words appearing close to each other and the order in which words appear as signals for a model to learn the meaning of words. The principle underlying the technique is that "a word is characterized by the company it keeps".[42] There is much technical evidence to show that vector embeddings representing words, which are often used as inputs to current state-of-the-art natural language processing systems, encapsulate gender biases.[43] An extensive study[44] looked into how stereotypical associations between gender and professional occupations propagate from the text used to train the models to the text embeddings, so that words like "doctor" are closely associated with the male gender pronoun "he".[45]

In the use of deep neural networks for supervised learning, engineers sometimes face the practical problem of having insufficient labelled data in their datasets. This is especially the case in applications where it takes domain experts to label the data, so that the creation of a huge, labelled dataset is a costly endeavour. To overcome the problem of limited training data, machine learning engineers often use a technique called transfer learning. This technique involves using a model already trained on another (possibly larger) dataset which contains data similar to the data the engineer is working with, before continuing training on the limited labelled data. Open-source models which have been pretrained on open datasets are

[41] Common Crawl is a US based non-profit organisation that systematically browses the world wide web and collects and stores data that are publicly available, available at https://commoncrawl.org (last accessed 15 June 2022).

[42] 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.

[43] C. Basta et al., "Extensive Study on the Underlying Gender Bias in Contextualized Word Embeddings", 33 (2021) Neural Computing and Applications 3371.

[44] Ibid.

[45] 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.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

made widely available by universities and technology companies. However, the geographic distribution of images in the popular ImageNet dataset reveals that 53 per cent 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, such as Open Images.[46] 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, there is another notion of fairness more generally applicable to all machine learning models that might not be clearly addressed by existing laws. This notion, which is considered in the machine learning literature on fairness, relates to the disparities in the performance of machine learning models with respect to data related to different demographic groups. These disparities can occur, for instance, when such 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.[47] Sentiment analysis tools, which predict the sentiment expressed by texts through assigning scores on a scale, have been shown to systematically assign different scores to text based on race-related or gender-related names of people mentioned.[48] Moreover, annotators' insensitivity to differences in dialect has also resulted in automatic hate speech detection models displaying a racial bias, so 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.[49] Even compared to human graders who may themselves give biased ratings,

[46] N. Meharbi et al., "A Survey on Bias and Fairness in Machine Learning" (2019) arXiv preprint arXiv:1908.09635.

[47] R. Tatman, "Gender and Dialect Bias in YouTube's Automatic Captions", in First ACL Workshop on Ethics in Natural Language 2017, 53–59.

[48] S. Kiritchenko and S.M. Mohammad, "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems" (2018) arXiv Preprint arXiv:1805.04508.

[49] 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.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

automated essay grading systems tend to assign lower scores to some demographic groups in a systemic manner.[50]

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".[51] 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 such as image classification, object detection and facial recognition. However, as previously discussed,[52] populations outside the US and UK are underrepresented in the standard datasets used for training such models. These datasets, curated predominantly 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.[53] It has also been found that 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 per cent difference in error rate.[54] 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.[55] First, the informational needs of some searchers or users may be served better than those of others. Harm 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

[50] 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.

[51] Barocas et al., *Fairness and Machine Learning*.

[52] See Section III(A)(2) above. See also Meharbi et al., "Survey on Bias and Fairness".

[53] T. de Vries et al., "Does Object Recognition Work for Everyone?" in 2019 ICCV 52–59.

[54] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification" (2018) 81 Proceedings of Machine Learning Research 1.

[55] Barocas et al., *Fairness and Machine Learning*.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

the relevant target variable of satisfaction is hard to measure:[56] clicks and ratings only serve as crude proxies for 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 suppressed by YouTube algorithms, while allegations relating to search have included partisan bias in search results.[57] Third, representational harms can occur by the amplification and propagation of cultural stereotypes.

4. Risk assessment tools

In risk assessment tools like COMPAS,[58] calibration[59] is an important goal. Equalised calibration requires that "outcomes are independent of protected characteristic after controlling for estimated risk".[60] For example, in a group of loan applicants estimated to have a 20 per cent 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.[61]

IV. Competing Algorithmic Fairness Metrics and Trade-offs

A. Fairness Metrics of Supervised Classification Models

Although the concept of fairness[62] in the law governing data processing is nebulous, the technical machine learning community has developed several

[56] Ibid.

[57] Ibid.

[58] Correctional Offender Management Profiling for Alternative Sanctions. It assesses the risk of recidivism for offenders.

[59] 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" in I. Guyon et al. (eds.), *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, available at https://papers.nips.cc/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html (last accessed 15 June 2022).

[60] 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.

[61] Pleiss et al., "On Fairness and Calibration".

[62] 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 are processed. Such 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 centred around having equal outcomes for different demographic groups:

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

technical metrics of fairness. In this section, we attempt to give a flavour of the various main categories of technical fairness metrics.

To begin with, "Fairness through Unawareness" is an approach to machine learning fairness where the model simply ignores special category data like race and gender, also known as 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 data categories which are correlated with the protected characteristic,[63] thus leading to indirect discrimination. A classic example of this would be the removal of the protected characteristic of race in a dataset, but the retention of another feature of the dataset focusing 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 unfeasible in some applications where it would, for example, require the removal of gender from facial images, or the removal of words relating to protected characteristics 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.[64] 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 predicts either a positive or negative class (e.g. whether a person is positive or negative for a disease).

see Department for Digital, Culture, Media and Sport, "Data: A New Direction", 26–31, available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1022315/Data_Reform_Consultation_Document__Accessible_.pdf (last accessed 15 June 2022). 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 behaviour). 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.

[63] D. Pedreshi, S. Ruggieri and F. Turini, "Discrimination-aware Data Mining" (2008) KDD '08 560.

[64] These four fairness metrics are group fairness metrics which make comparisons between demographic groups. Another approach to fairness is that of individual fairness which involves looking 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 humans 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?" (2021), available at https://ssrn.com/abstract=3819799 (last accessed 13 November 2021).

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

1. *Demographic Parity*

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 which are differentiated based on the protected characteristic.[65] Between two demographic groups which are differentiated based on the race protected characteristic, namely Whites and African Americans, perfect satisfaction of this metric in 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[66] 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 (those who did actually repay the loan) and unqualified applicants (those who 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 the problems with Demographic Parity, an alternative metric called Equality of Odds was proposed. This metric computes both the difference between the false positive rates,[67] and the difference between the true positive rates,[68] of the decisions of the model on the two demographic groups across the protected characteristic.[69] For instance, enforcing this metric in relation to a model in our previous example would ensure that the rate of qualified African Americans getting a loan is equal to that of

[65] M. Hardt et al., "Equality of Opportunity in Supervised Learning" (2016) N.I.P.S., available at https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf (last accessed 15 June 2022).

[66] C. Dwork, et al., "Fairness through Awareness" (2012) I.C.T.S. '12 214.

[67] The fraction of negative cases which were incorrectly predicted to be in the positive class out of all actual negative cases: see S. Verma and J. Rubin, "Fairness Definitions Explained" (2018) FairWare '18 1.

[68] The fraction of positive cases which were correctly predicted to be in the positive class out of all actual positive cases: see ibid.

[69] Hardt et al., "Equality of Opportunity in Supervised Learning".

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

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 operation of the controversial COMPAS[70] 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 was twice that for African Americans than for Whites.[71] 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 reoffend, it was much more likely for the Whites to be predicted by the algorithm not to 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.[72] 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[73] to achieve a higher accuracy rate due to the lack of the additional constraint.

However, it has also been found that enforcing equality only in relation to the true positive rate will increase disparity between the demographic groups in relation to the false positive rate.[74] 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 reoffending among African Americans and Whites 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 African Americans who do not actually reoffend, would almost always result in a drop in the overall accuracy of the model. This is because in

[70] Correctional Offender Management Profiling for Alternative Sanctions. It assesses the risk of recidivism for offenders.

[71] J. Larson et al., "How We Analyzed the COMPAS Recidivism Algorithm", *ProPublica*, 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, "A Review on Fairness in Machine Learning" (2023) 55 A.C.M. Computing Surveys 1.

[72] Hardt et al., "Equality of Opportunity in Supervised Learning".

[73] Ibid.

[74] Pleiss et al., "On Fairness and Calibration".

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

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 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, risk scores are also provided by the model 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 African American 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.[75] If a classifier has equal false negative rates between both groups, it can be mathematically derived that it will also have equal

[75] A. Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments" 2016 (5) Big Data 153; J. Kleinberg, S. Mullainathan and M. Raghavan, "Inherent

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

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 the 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 (so that, for example, 80 per cent of people assigned a 0.8 risk score actually default), while also having (1) the rate at which individuals predicted to default do not actually default (the false positive rate) to be equal between both demographic groups and (2) the rate at which individuals predicted to not default actually default (the false negative rate) to be equal between both groups.

Further, it was found that, on the specific recidivism dataset on which COMPAS was used, enforcing an algorithm to achieve calibration would result in disparities in both the false positive and false negative rates[76] across demographic groups. On the other hand, mis-calibrated risk scores would cause discrimination to one of the demographic groups, since a user of the model would need to interpret the risk scores differently depending on the demographic group 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.[77]

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 either an equalised false positive rate or an equalised false negative rate, corresponding to the condition for which consequences are more dire.[78] 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 which bears 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

Trade-offs in the Fair Determination of Risk Scores" (2017) 67 Innovations in Theoretical Computer Science 1.

[76] Pleiss et al., "On Fairness and Calibration".

[77] Ibid.

[78] Kleinberg et al., "Inherent Trade-offs in the Fair Determination of Risk Scores".

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

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 an 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 mean that the algorithm recommended that the company pay both of them more than $50,000,[79] even if one of them does not deserve 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.[80] 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.

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 models 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: (1) disclosure of the machine learning models used; (2) disclosure of the fairness metrics and the trade-offs involved; (3) disclosure of any de-biasing methods adopted; and (4) release of datasets for public inspection or for third-party audit.

[79] Pleiss et al., "On Fairness and Calibration".

[80] R. Berk et al., "Fairness in Criminal Justice Risk Assessments: The State of the Art" (2021) 50 Sociological Methods and Research 3.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

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 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, crowd-sourcing 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.[81] 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 in relation to such trade-offs and the reasons behind it.

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 such as monetary rewards,[82] 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 forward with concerns over the environmental and social impacts of the

[81] Barocas et al., *Fairness and Machine Learning*.

[82] For example, the US Securities and Exchange Commission (SEC) will pay a monetary award to whistleblowers 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, Form WB-APP, available at https://www.sec.gov/files/formwb-app.pdf (last accessed 15 June 2022).

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

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 were used in recent significant enhancements to Google's core search product,[83] more than 1,400 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 obtained from 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 such matters to light without suffering personal repercussions.

To address disparities in the performance of models, 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 should consider opening up a limited interface for non-commercial use of their AI applications, where public users can probe the model to check for fairness. For example, registered users could 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 De-biasing 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 de-biasing 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 shown to

[83] K. Hao, "We Read the Paper that Forced Timnit Gebru out of Google: Here's What it Says", *MIT Technology Review*, available at https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/ (last accessed 15 June 2022).

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

have the actual effect of mostly hiding the bias, not removing it.[84] The gender bias information can still be found in the vector space distances between "gender-neutralised" words in the de-biased vector embeddings and is still recoverable from them.[85] This is why the techniques for de-biasing have to be reported in conjunction with the fairness metrics: to prevent companies "over-optimising" on the chosen fairness metric in the way described above, without serving the actual goal of fairness. It is important to note that the de-biasing techniques used can be reported with little to no revelation about the AI model itself. Thus, companies should have no excuse for not reporting 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 re-labelling 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 be likely to 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",[86] but it remains to be seen whether such methods will result in other trade-offs.

Because deep learning models learn rich representations of the data with which they are fed, deep learning researchers have experimented to see if models can learn fair representations.[87] For example, representations of the data learnt by the deep learning model, instead of the raw data itself, are used to make predictions. If the way the representations of the data are learnt is constrained in a manner that excludes information on demographic group membership, then the predictive part of the model has no

[84] H. Gonen and Y. Goldberg, "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings but Do Not Remove Them" (2019 N.A.A.A.C.L. arXiv:1903.03862.

[85] The study mentioned in the previous footnote 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 was not determinative of gender bias in the word embedding. Thus, even though methods of de-biasing 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.

[86] J.E. Johndrow and K. Lum, "Algorithm for Removing Sensitive Information: Application to Race-independent Recidivism Prediction" (2017).

[87] H. Zhao and G.J. Gordon, "Inherent Tradeoffs in Learning Fair Representations" in H. Wallach et al. (eds.), *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, available at https://papers.nips.cc/paper/2019/file/b4189d9de0fb2b9cce090bd1a15e3420-Paper.pdf (last accessed 15 June 2022).

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

discernible information about group membership to work with in making its predictions. Thus, the decisions would be made in a manner independent of group membership, which is what researchers who work on fair representations argue is a fairer approach.[88]

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.[89]

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.[90] 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 de-biasing techniques is still nascent, though there is evidence of consequences for both model accuracy and trade-offs, with competing fairness goals not taken into account by the choice of de-biasing technique. Making it mandatory for companies to transparently report any de-biasing 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.[91] However, it is understandable that significant investment is often required on the part of companies to

[88] However, learning fair representations presents a trade-off between Demographic Parity (see Section IV (A)(1) above) and the accuracy of the models trained, when the phenomenon to be predicted occurs at different rates in reality between the different demographic groups in question. See Zhao and Gordon, "Inherent Tradeoffs in Learning Fair Representations".

[89] T. Kamishima et al., "Fairness-aware Learning Through Regularization Approach" (2011) 2011 IEEE 11th International Conference on Data Mining Workshops 643.

[90] Hardt et al., "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.

[91] 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.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

collect and curate such datasets in order to obtain a competitive advantage. As a result, companies might be reluctant to share this data. Also, some datasets might also contain trade secrets, confidential information and the private data of users. It might not always be feasible to completely prevent data re-identification from the release of anonymised data. Thus, the release of datasets should not be mandated, but is best left to a comply-or-explain basis.[92]

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 licence.[93] A new standard data release licence, similar to non-commercial and no derivatives licences used for research data,[94] can be created in such a way that the use of the data is limited to inspection for fairness concerns. Admittedly, enforcement of such a licence can be a problem if it is possible for models to be trained using 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 about 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[95] to values in the dataset have had to be made to prevent the 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

[92] MacNeil and Esser, "The Emergence of 'Comply or Explain'".

[93] "Project Open Data", available at https://project-open-data.cio.gov/open-licenses/ (last accessed 13 November 2021).

[94] "OpenAire", available at https://www.openaire.eu/research-data-how-to-license/ (last accessed 13 November 2021).

[95] 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.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

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.

VI. Application of AI Fairness Reporting Framework to Two Case Studies

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

We consider the case of Goldman Sachs' credit profiling of applicants for the Apple Card. A technology entrepreneur, David Heinemeier Hansson, raised concerns about Goldman Sachs' Apple Card program for gender-based discrimination through the use of what he called a "black-box algorithm".[96] He claimed that, although he and his wife filed joint tax returns and lived in a community-property state, he received a credit limit that was 20 times higher than that offered to his wife. Hansson also expressed concerns that "the Bank relied on algorithms and machine learning for credit decision-making, and [he] complained that an Apple Card customer service agent could not explain these algorithms or the basis for the difference in credit limits".[97] Apple's co-founder 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 retrospectively 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[98] by the New York State Department of Financial Services report could have at least been mitigated, if not avoided entirely.

1. Disclosing all uses of machine learning models

Under the proposed AI Fairness Reporting framework, Goldman Sachs would have needed to disclose all its uses of machine learning models as a matter of best practice. Disclosure of even the use of machine learning

[96] S. Perez, "New York's Department of Financial Services Says Apple Card Program Didn't Violate Fair Lending Laws", *Techcrunch*, available at https://tcrn.ch/3sjjlOD (last accessed 15 June 2022).

[97] New York State Department of Financial Services, "Report on Apple Card Investigation" (March 2021), 4, available at https://www.dfs.ny.gov/system/files/documents/2021/03/rpt_202103_apple_card_investigation.pdf (last accessed 15 June 2022).

[98] Ibid.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

models which have not been making directions or predictions directly affecting individuals would have been needed under our reporting framework. This would have included internal risk management models which predicted the health of Goldman Sachs' lending business. If the internal risk models had consistently predicted a high-risk exposure to Goldman Sachs' lending business just before 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 have raised fairness considerations.

The machine learning models used in Goldman Sachs relating to the Apple Card program, which directly affected individuals, included more than just the credit scoring model. Under our proposed reporting framework, machine learning models deployed on Goldman Sachs' consumer-facing platforms, which determined whether to advertise or recommend the Apple Card to a particular user, would have been needed to go through detailed fairness reporting as well.

2. Reporting on fairness metrics used

Under our proposed reporting framework, the choice of fairness metrics should have taken 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. Under these laws, the gender of credit applicants cannot be taken into account in the credit decisions and two categories of discrimination are recognised: disparate treatment and disparate impact. Under our proposed reporting framework, de-biasing a machine learning model, together with the disclosure of group fairness metrics, would have revealed that protected characteristics like gender had been taken into account. If so, this would have contravened the disparate treatment requirement since the Equal Credit Opportunity Act 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 degree of independence between gender and the credit decisions made.[99] The Department

[99] 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.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

found that the Bank had a fair lending programme 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".[100] The New York State Department of Financial Services, in its investigation report,[101] 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, under our proposed reporting framework Goldman Sachs would have had to choose both a group fairness metric and an individual fairness metric to report on.[102] 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 have resulted in individuals with highly similar profiles being given a different credit rating. This could have happened 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 have been assigned different outcomes. All metrics used in arriving at the operational model should have thus been reported to show transparently how these trade-offs were navigated in the final model used.

3. Reporting on de-biasing methods used

What is completely missing in both the investigation report and subsequent public relations efforts by Goldman Sachs on the Apple Card program is an account of any specific de-biasing methods used to arrive at the fairness outcomes, which we propose should have been made public.

Existing laws like the Equal Credit Opportunity Act serve to protect consumers from discriminatory lending based on protected characteristics, so the investigation report's finding that no fair lending laws have been breached serves little to inform other stakeholders on how the use of the machine learning model affects them. Investors and stakeholders of Goldman Sachs would have been interested to know how much the de-biasing methods used (if any) would have had an impact on the accuracy of the credit scoring model as this would have affected the business and operations of Goldman Sachs, which would have in turn impacted its financial performance and reputation. Researchers could have further

[100] 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.

[101] New York State Department of Financial Services, "Report on Apple Card Investigation", 4.

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

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

concentrated their study of the implications of such de-biasing techniques being used in practice, in the specific context of credit scoring, given that the full implications of de-biasing techniques are still under-researched. Credit applicants themselves would have wanted to know how such de-biasing techniques might have potentially affected them and therefore would have wanted a fuller report that did not merely confirm that there was compliance with the law.

4. Release of datasets for inspection

We refer here to the German Credit Dataset[103] 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 1,000 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.

Under our proposed reporting framework, 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 that data.

However, even if Goldman Sachs had 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 were 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[104] in the Department's report on the Apple Card case as to why spouses with shared bank accounts and assets 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[105] – the applicant would not know the attributes that Goldman Sachs took into account in its credit scoring model.

[103] A standard credit scoring dataset used in machine learning fairness research. Available at https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data) (last accessed 15 June 2022).

[104] New York State Department of Financial Services, "Report on Apple Card Investigation", 10.

[105] Ibid., at 9. The "A Closer Look at our Application Process" portion of the website provides a snapshot of the data the Bank draws upon in setting credit terms.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

B. Wrongful Arrest Attributed to False Positive Match by the 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 the Dataworks Plus facial recognition software.[106]

This culminated in a request by Senator Sherrod Brown of Ohio for Dataworks Plus to provide information to the US Congress on questions including (1) whether the company planned to impose a moratorium on the use of its facial recognition technologies by law enforcement, (2) what the factual basis behind marketing claims by the company on the reliability and accuracy of its facial recognition system was and (3) whether there was an executive responsible in the company for facilitating conversations on ethical decision-making.[107] Keeping in mind that Dataworks Plus brands itself as a "leader in law enforcement and criminal justice technology,"[108] with the facial recognition system FACE Plus being one of its key offerings, imposing such a moratorium would have a substantial impact on its financial revenue.

This case is different from the previous case[109] in that the creator of the facial recognition system was not the user of the system: that was the Detroit police department. Also, there is a nuanced difference here in relation to the allegation of unfairness. This was 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 matched facial snapshots from crime scene video surveillance to a 50 million Michigan police database of driver's licence photographs in order to generate matches to candidates who might be potential suspects. The allegation was that the quality of matches 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 on such systems[110] it was found that the error rate for light-skinned men is never worse than 0.8 per cent, but 34.7 per cent for dark-skinned women. According to the study, although researchers at a major US technology company claimed

[106] K. Hill, "Wrongfully Accused by an Algorithm", *New York Times*, available at https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html (last accessed 15 June 2022).

[107] Ibid.

[108] "Dataworks Plus", available at http://www.dataworksplus.com/index.html (last accessed 13 November 2021).

[109] See Section VI(A) where Goldman Sachs was the creator and user of the credit profiling system.

[110] L. Hardesty, "Study Finds Gender and Skin-type Bias in Commercial Artificial-intelligence Systems", *MIT News*, available at https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 (last accessed 15 June 2022).

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

an accuracy rate of more than 97 per cent for a face-recognition system they had designed, the dataset used to assess its performance was more than 77 per cent male and more than 83 per cent White. A National Institute of Standards and Technology study[111] covered 189 software algorithms from 99 developers, which make up the majority of the industry in the US. The study 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[112] which are commonly used in suspect identification systems, there was a higher rate of false positives for African American women, although the study contained a caveat pointing out that not all algorithms give this high rate of false positives across demographics in these types of system and systems that are the most equitable are also amongst the most accurate. By the account of the Detroit Police Chief, the Dataworks Plus facial recognition system misidentifies 96 per cent of the time.[113] 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 to Dataworks Plus, we argue that the process would have enabled Dataworks Plus to identify problems better with its facial recognition system and would have allowed the civilian oversight board[114] in Detroit to evaluate the adoption of the system better. The discussion in Sections 1 to 4 below describe the consequences of applying the requirements of our proposed reporting framework to the facts of the Dataworks Plus case.

1. Disclosing all uses of machine learning models

Under our proposed AI Fairness Reporting framework, Dataworks Plus, being a provider of software systems rather than a user, would have needed to disclose all the uses of machine learning models in the various software solutions it provided. There might have been 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.

[111] P. Grother et al., "Face Recognition Vendor Test" (FRVT) Part 3: Demographic Effects" (2019) Internal Report 8280, available at https://nvlpubs.nist.gov/nistpubs/ir/2019/nist.ir.8280.pdf (last accessed 15 June 2022).

[112] Systems which generate matches to multiple candidates as suspects from a single photograph, rather than matching to a single candidate.

[113] J. Koebler, "Detroit Police Chief: Facial Recognition Software Misidentifies 96% of the Time", *Vice*, available at https://www.vice.com/en/article/dyzykz/detroit-police-chief-facial-recognition-software-misidentifies-96-of-the-time (last accessed 15 June 2022).

[114] "Board of Police Commissioners", available at https://detroitmi.gov/government/boards/board-police-commissioners (last accessed 15 June 2022).

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

We do note that there might have been concerns about the protection of trade secrets, if the disclosure of machine learning model use were made compulsory. However, there could have been a degree of flexibility afforded to the company with regards to the granularity of disclosure: the disclosure could have ranged from the general class of machine learning model to the specific model used. It would have been hard for a company to justify why such a requirement, modified by the flexibility mentioned above, could not have been imposed on companies; especially when it is balanced against the interests of stakeholders such as 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[115] studied the differences in false positives and false negatives between demographic groups in the dataset, along the lines of gender, age and racial background. We suggest that these two metrics would have been apt for use in AI Fairness Reporting by Dataworks Plus. This would have been a holistic representation of how well the facial recognition system had performed, in stark contrast to the marketing materials on the Dataworks Plus website that were highlighted by Senator Brown, which vaguely described the identification of the facial candidates produced by the FACE Plus software system as "accurate and reliable".

When the wrongful arrest of Robert Julian-Borchak Williams, mentioned earlier, was first reported in the New York Times, the General Manager of Dataworks Plus, Todd Pastorini, was cited as claiming that checks which Dataworks Plus did when they integrated facial recognition systems from subcontractors were not "scientific" and that no formal measures of the systems' accuracy or bias were done. All this negative publicity 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 included components developed by two other companies, NEC and Rank One Computing.[116] The NIST study[117] 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.[118]

However, one more nuance needs to be appreciated in this situation where the developer of the AI system was not the end user: the prediction

[115] Grother et al., "Face Recognition Vendor Test".
[116] Hill, "Wrongfully Accused by an Algorithm".
[117] Grother et al., "Face Recognition Vendor Test".
[118] Hill, "Wrongfully Accused by an Algorithm".

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

outputs of the AI system needed to be interpreted and acted upon by the users who were not as familiar as developers with the workings of machine learning models. In the Dataworks 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.[119] It was up to the investigator to interpret these matching candidates, along with the associated confidence scores, before deciding whether to proceed with any arrest. The outputs of the AI system were thus characterised by law enforcement and software providers like Dataworks Plus as mere investigative leads and were therefore not conclusive as to arrest decisions. In such a situation, assuming proper use of the system, the presence of false positives was 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 have been helpful if included in the AI Fairness Reporting.

3. Reporting on de-biasing methods used

The Dataworks Plus case presented a clear risk that the use of de-biasing methods could have created other problems. A study[120] 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 for White people, unless they are 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 were used, which means it would make it easier for the actual White culprits to evade detection by the system. Using different models, however, would have required a separate classification model for choosing the appropriate model to use, or have required the police to exercise judgment which might introduce human bias.[121] It is, therefore, important that the methods used to address bias were disclosed in order that observers could anticipate and flag any potentially inadvertent problems that the models created.

4. Release of datasets for inspection

The datasets contained the photographs of individuals, which made anonymisation without removing important information in the data practically

[119] Ibid.

[120] A. Harmon, "As Cameras Track Detroit's Residents, a Debate Ensues Over Racial Bias", *New York Times*, available at https://www.nytimes.com/2019/07/08/us/detroit-facial-recognition-cameras.html (last accessed 15 June 2022); K.S. Krishnapriya et al., "Characterizing the Variability in Face Recognition Accuracy Relative to Race" in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (Long Beach 2019), 2278.

[121] Hill, "Wrongfully Accused by an Algorithm".

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

impossible. However, under our proposed AI Fairness Reporting framework, the metadata of the subjects could have been released and reference could have been made to the metadata information used in the NIST study[122] indicating the subject's age, gender and either race or country of birth. This transparency with regards to metadata information would have allowed underrepresentation of demographic groups in the dataset to be detected and flagged by observers and in our view would have been 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 use of AI, including between AI fairness metrics as well as accuracy. There are also reported incidents of bias in AI systems which have captured the public consciousness, leading to a backlash against companies in the form of employee walkouts, resignations of key executives[123] and media scrutiny.[124]

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, such as 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.

[122] US Department of Commerce, National Institute of Standards and Technology, "NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software", available at https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software (last accessed 15 June 2022).

[123] J. Dastin and P. Dave, "Google AI Scientist Bengio Resigns After Colleagues' Firings: Email", *Reuters*, available at https://www.reuters.com/business/media-telecom/google-ai-scientist-bengio-resigns-after-colleagues-firings-bloomberg-2021-04-06/ (last accessed 15 June 2022).

[124] R. Mac, "Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men", *New York Times*, available at https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html (last accessed 15 June 2022).

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press

Companies might also push back on how the AI Fairness Reporting would create an onerous burden for companies using AI and would hold the use of AI to a higher standard of interrogation than that applied to human decision makers. However, it is important to note the opportunity opened up by the use of AI for unfairness to be combated, which was not available with human decision makers. Despite the complaints about the opacity of AI, AI would still be 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. 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 de-biasing 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,[125] in the hope that efforts to ensure companies act more responsibly towards their stakeholders, such as the proposed AI Fairness Reporting, can have similar traction.

[125] The UK Companies (Strategic Report) (Climate-related Financial Disclosure) Regulations 2022 amended certain sections in the Companies Act 2006 to require certain publicly quoted companies and large private companies to make disclosures aligned with the Financial Stability Board's Task Force on Climate-related Financial Disclosures in their annual reports.

https://doi.org/10.1017/S0008197322000460 Published online by Cambridge University Press