ALGORITHMIC FAIRNESS

Challenges and Opportunities for Artificial Intelligence Governance

Across the world, artificial intelligence ("AI") is increasingly used to automate decision making for access to pivotal socio-economic opportunities such as university applications and credit ratings. Given recent high-profile examples of algorithmic bias in those areas, it is timely to consider what it means for AI to be fair and how to achieve that. This article spotlights the growing field of algorithmic fairness, which seeks to directly incorporate fairness into AI algorithms, and identifies two key findings that have significant implications on AI governance. The authors also examine whether existing legal and regulatory frameworks in Singapore, the EU, the US and China can adequately regulate AI decision making, especially in light of the two important findings. The article concludes with several recommendations for how the algorithmic fairness field can further contribute to the ongoing development of AI governance.

Shaun **KHOO**[1]
BA (Oxford), MA (Columbia University);
Data Scientist, Government Technology Agency Singapore.

CHOW Zi En
LLB (Singapore Management University), LLM (New York University);
State Counsel, Attorney-General's Chambers Singapore.

I. Introduction

1 From ordinary tasks like typing emails[2] to creative exploits such as composing music,[3] artificial intelligence ("AI") is now embedded in our lives in ways few could have imagined a few decades ago. Unsurprisingly,

1 The views expressed in this article are those of the authors and are not representative of the views of their employers.

2 Andrew Dai *et al*, "Gmail Smart Compose: Real-Time Assisted Writing" *Google Research* (25 July 2019) <https://dl.acm.org/doi/abs/10.1145/3292500.3330723> (accessed 15 March 2022).

3 Maura Barrett & Jacob Ward, "AI can now compose pop music and even symphonies. Here's how composers are joining in" *NBC News* (30 May 2019) <https://www.

(cont'd on the next page)

AI has also been used to automate decisions that previously relied entirely on human discretion, such as pricing strategy, recruitment, credit scoring, fraud detection and even parole decisions.

2 The proliferation of AI comes with its set of challenges. In 2012, the New York Times revealed how a department store was predicting which shoppers were likely to be pregnant based on their purchase histories. In an iconic example of AI-driven marketing, the father of a high school girl was outraged when the store sent her coupons for baby items, only to return embarrassed a few days later when he discovered his daughter was actually pregnant as predicted.[4] In more sensitive areas like finance and medicine, the consequences go beyond awkwardness. In 2019, David Heinemeier Hansson applied for an Apple Card and received a credit limit 20 times higher than his wife's, despite filing joint tax returns and co-owning the property they lived in.[5] AI models have also been found to significantly underestimate the healthcare needs of black patients relative to white patients, because black patients historically incurred lower medical costs than white patients despite having the same number of health issues.[6]

3 Some kind of governance is no doubt overdue. However, figuring out the details is a tall order. Unlike traditional computer algorithms, AI models are not hard-coded with rules, but instead infer patterns from data to generate predictions. These models can be extremely complex – GPT-3, one of the largest AI models to date, has 175 billion parameters.[7] As such, how AI models arrive at their decisions are often inscrutable even to the data scientists who developed them. Moreover, bias can seep into AI models in many ways, from measurement biases in how the data was collected, to the choices of what features to include in the model. Further, companies often refuse to share or release their models publicly as the AI models are treated as trade secrets. These factors make it hard for AI governance to develop.

---

nbcnews.com/mach/science/ai-can-now-compose-pop-music-even-symphonies-here-s-ncna1010931> (accessed 15 March 2022).

4 Charles Duhigg, "How Companies Learn Your Secrets" *The New York Times* (16 February 2012) <https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html> (accessed 10 April 2022).

5 Neil Vigdor, "Apple Card Investigated After Gender Discrimination Complaints" *The New York Times* (10 November 2019) <https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html> (accessed 10 February 2022).

6 Carolyn Y Johnson, "Racial bias in a medical algorithm favors white patients over sicker black patients" *The Washington Post* (24 October 2019) <https://www.washingtonpost.com/health/2019/10/24/racial-bias-medical-algorithm-favors-white-patients-over-sicker-black-patients/> (accessed 10 April 2022).

7 Tom B Brown *et al*, "Language Models are Few-Shot Learners" *OpenAI* (22 July 2022) <https://arxiv.org/pdf/2005.14165.pdf> (accessed 15 March 2022).

4 Current discussions on AI governance typically focus on the organisational structure and processes around the AI model, such as setting up a code of ethics for the use of AI and developing risk management frameworks for AI applications.[8] This article aims to contribute to the ongoing discourse on AI governance by examining how algorithmic fairness, a field concerned with developing algorithmic approaches to promoting fairness, can be a critical part of AI governance. In particular, several key findings from the field have important implications on how the risks of AI bias can be managed.

5 Part II provides a detailed account of the concept of algorithmic fairness. Part III discusses two seminal findings within the algorithmic fairness field and describes the implications this has on AI governance. Part IV explores the current legal and regulatory landscape for AI models, focusing specifically on aspects relevant to the findings in Part III. Part V articulates three proposals for AI governance. Part VI concludes the article.

A. Defining artificial intelligence

6 As a rapidly growing space, many terms relating to machine learning and AI are constantly being refined and used in a myriad of contexts. To avoid ambiguity about the subsequent arguments, this article sets out the definition of AI below.

7 This article borrows the definition of AI from Singapore's Model AI Governance Framework ("Model Framework"), where AI "refers to a set of technologies that seek to simulate human traits such as knowledge, reasoning, problem solving, perception, learning and planning, and, depending on the AI model, produce an output or decision (such as a prediction, recommendation, and/or classification). AI technologies rely on AI algorithms to generate models. The most appropriate model(s) is (or are) selected and deployed in a production system".[9] Note that this article regards machine learning as being interchangeable with AI.

---

8 "Model Artificial Intelligence Governance Framework: Second Edition" *Personal Data Protection Commission Singapore* (21 January 2020) <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf> (accessed 3 January 2022).

9 "Model Artificial Intelligence Governance Framework: Second Edition" *Personal Data Protection Commission Singapore* (21 January 2020) <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf> (accessed 3 January 2022).

8 The distinction between AI algorithms and AI models is a subtle but important one.[10] AI algorithms produce AI models which optimise a specified performance metric given a fixed dataset. AI models are just deterministic mathematical equations where the input data is computed using the trained (or finalised) parameters to generate the predictions, while AI algorithms are the optimisation algorithms that seek to identify the top AI model (or set of parameters) which best satisfies the performance metric. A helpful analogy is the distinction between a blueprint and a product. AI algorithms are the blueprints which set out how to build the product, while AI models represent the product itself. For example, the artificial neural network is a well-known AI algorithm, while the specific AI model, trained using that algorithm, is the object actually integrated into Google Search.[11] Extending the comparison, both the blueprint and the materials matter to the quality of the final product. If the blueprint was not designed with safety in mind, then harmful incidents occurring is inevitable. Analogously, if AI algorithms were not designed with fairness in mind, then algorithmic bias should come as no surprise. If the materials are of poor quality, then the final product is likely to be substandard. Similarly, missing or biased data results in inaccurate or biased AI models. This conceptual distinction will be used in the rest of the paper to diagnose and analyse the exact source of bias in AI systems.

II. Introducing algorithmic fairness

9 This Part provides context to and defines the concept of algorithmic fairness, specifically how biases in AI come about, why fairness considerations in AI are so important and unique, and what algorithmic fairness is.

A. Bias in artificial intelligence

10 Although AI is often touted as a force for good, recent history is replete with examples where AI has perpetuated bias. In the recruitment and hiring space, AI was initially advocated as a solution to slow and biased hiring processes, since AI models "[rely] on data to surface candidates from a wide variety of places and match their skills to the

10 Kleinberg *et al* draw a similar distinction between the "trainer algorithm" (the AI algorithm) and the "screener algorithm" (the AI model): See Jon Kleinberg *et al*, "Discrimination in the Age of Algorithms" (2018) 10 *Journal of Legal Analysis* 113 at 115.

11 Pandu Nayak, "Understanding Searches Better than Ever Before" *Google* (25 October 2019) <https://blog.google/products/search/search-language-understanding-bert/> (accessed 15 March 2022).

job requirements, free of human biases".[12] Reality suggests otherwise. In 2014, Amazon developed an AI model to score job applicants based on their resumes, but soon noticed that the algorithm was biased against female applicants. Notably, the AI model penalised applicants whose resumes contained the word "women's" and downgraded the scores of applicants from two all-women's colleges. Amazon ultimately dropped the whole effort as it could not completely debias the model.

11 Recalling the earlier distinction between AI algorithms and AI models, while AI models are the engines perpetuating bias, they are not the source of biases themselves. It is the process of developing these AI models that deserves greater attention. As Kleinberg *et al* put it, "the Achilles' heel of all algorithms is the humans who build them and the choices they make".[13] This prompts the search for an algorithmic approach to ensuring fairness: by designing fairness into the AI algorithm directly, the risk of human biases entering AI models can be directly mitigated.

B. Fairness in artificial intelligence

12 Fairness is generally considered as an essentially contested concept, where there is widespread agreement on its importance but not on how it should be actualised.[14] In the context of AI-driven decision making, fairness takes on a more specific and tangible definition – ensuring that AI models "[treat] similar individuals similarly".[15] Two people who are equally qualified for a job should receive the same predictions from the AI model. Irrelevant factors, like one's height or race when applying for a software programmer role, should have no influence over the prediction.

13 How are concerns about fairness different for algorithms than for humans? While there are some commonalities, there are two important distinctions that warrant particular attention on fairness in AI. First, algorithms may be more transparent than humans in some ways. Algorithmic decision making has been routinely criticised for being a black box, and this is true for complex AI models. However, AI

12 Claire Cain Miller, "Can an Algorithm Hire Better Than a Human?" *The New York Times* (25 June 2015) <https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html> (accessed 20 December 2021).

13 Jon Kleinberg *et al*, "Discrimination in the Age of Algorithms" (2018) 10 *Journal of Legal Analysis* 113 at 117.

14 Walter B Gallie, "Essentially Contested Concepts" (1956) 56 *Proceedings of the Aristotelian Society* 167.

15 Cynthia Dwork *et al*, "Fairness Through Awareness" *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (8 January 2012) <https://arxiv.org/pdf/1104.3913.pdf> (accessed 14 April 2022).

algorithms must be instructed what to predict, what data points to use, and what to optimise for, while AI models are deterministic mathematical equations which convert data points into a prediction. These explicit choices made by the data scientists will be encoded within the AI model and its documentation, and can be easily opened up for greater scrutiny. Moreover, there have been significant advances in unpacking black box models and making them more interpretable.[16] In contrast, human decision making is fraught with cognitive errors, implicit biases, and *post hoc* reasoning. Even well-intentioned people may have blind spots affecting their decisions.[17] In situations where quick decisions have to be made, such as whether to interview a particular candidate out of thousands that applied, biases can unwittingly seep in to corrode prudent decision making. Hence, AI offers society an opportunity to apply a more precise, transparent and consistent approach to important decisions. In some situations, AI can be fairer in both process and substance, especially when compared to human decision-makers who conceal or obfuscate their biases.

14 Second, algorithmic decision making happens at an unprecedented speed and scale. AI models take mere seconds to determine credit risk for a housing loan, while an experienced manager may need half an hour for a proper decision. Moreover, AI models can be rapidly scaled up to serve more locations, while human decision makers need to be trained extensively to make correct decisions consistently. As such, AI has considerably greater potential to perpetuate and amplify the harms arising from ingrained biases. With AI integrated into many aspects of our lives, it is essential to think carefully about how to design AI algorithms correctly to ensure they work as intended.

15 These two facts underscore the uniqueness and criticality of specifying fairness for AI correctly. While existing philosophical and legal work on fairness are still largely applicable, they must be contextualised to the new challenges that AI-driven decision making poses. Synthesising conceptual debates with robust mathematical foundations to weave fairness into the fabric of AI itself is what drives the field of algorithmic fairness.

16 Riccardo Guidotti *et al*, "Factual and Counterfactual Explanations for Black Box Decision Making" (2019) 34(6) IEEE Intelligent Systems 14.

17 Mahzarin R Banaji & Anthony G Greenwald, *Blindspot: Hidden Biases of Good People* (Delacorte Press, 2013).

C. Algorithmic fairness

16 The earliest papers in the field of algorithmic fairness emerged in the late 2000s. Pedreschi *et al* demonstrated how AI models can be biased against minority groups even if data on their minority status is withheld,[18] while the later work by Dwork *et al* in formulating a framework for algorithmic fairness[19] inspired many researchers to enter the field. Today, less than two decades later, algorithmic fairness is now a key part of many academic conferences for machine learning and AI.

17 Given the relative nascency of this field, there is no clear definition for algorithmic fairness as it is often used as a catch-all for work relating to fairness in the field of computer science. For the purposes of this article, algorithmic fairness refers to using algorithmic[20] approaches to ensure fairness in the application of AI. This spans a wide range of research, from debiasing datasets, measuring bias in large AI models, to designing fairness constraints into AI algorithms. This also encompasses a spectrum of AI applications, from enhancing existing technological tools (such as voice recognition) to AI-driven decision making (such as automated hiring).

18 As algorithmic fairness is a broad field, this article will focus on a narrower subset: harm caused by AI-driven decision making which affects individuals' access to important socio-economic opportunities, including housing loans, recruitment and grant applications, given the serious consequences on individuals' lives. This is known as allocative harm, which occurs "when a system allocates or withholds certain groups an opportunity or a resource".[21] A lot of work in algorithmic fairness, especially in earlier articles, have focused on allocative harms as these have serious consequences on one's life and often appear in the media as prominent cases of AI bias. As such, this article will adopt a similar focus.

---

18 Dino Pedreschi *et al*, "Discrimination-aware Data Mining" *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (24 August 2008) <http://pages.di.unipi.it/ruggieri/Papers/kdd2008.pdf> (accessed 14 April 2022).

19 Cynthia Dwork *et al*, "Fairness Through Awareness" *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (8 January 2012) <https://arxiv.org/pdf/1104.3913.pdf> (accessed 14 April 2022).

20 "Algorithm" generally refers to "a set of mathematical instructions or rules that, especially if given to a computer, will help to calculate an answer to a problem". Definition obtained from the *Cambridge Advanced Learner's Dictionary & Thesaurus* (Cambridge University Press).

21 Kate Crawford, "The Trouble with Bias" *YouTube* (5 December 2017) <https://www.youtube.com/watch?v=fMym_BKWQzk> (accessed 12 January 2022).

III. Making AI-driven decision making fairer

19 This Part introduces two key findings from algorithmic fairness that have significant implications on AI governance.

A. Fairness through awareness

20 One of the mainstream approaches to tackling discrimination is to prevent the collection of data on sensitive attributes, which are personal characteristics that identify specific groups like gender, race, religion or age. The justification is that without data on the applicant's sensitive attributes, it would be impossible to discriminate on those grounds.

21 However, this approach is far from perfect. Information about the sensitive attribute can be inferred from other data points provided. Gender is easily identifiable from activities that are highly correlated with gender. In Singapore's context, National Service is one such example. Age can be calculated by looking at when the individual graduated from university, a ubiquitous and reasonable requirement in job applications. Given how easy it is for human decision-makers, AI models unsurprisingly do the same. An audit done on Facebook's ad-delivery system found that the jobs which were traditionally male-dominated were mostly advertised to men, despite Facebook disabling the algorithm from targeting a specific gender.[22]

22 Furthermore, blinding the AI model from accessing the sensitive attribute may even lead to worse outcomes for disadvantaged groups. Using college grades data, researchers found that a race-aware predictive model was better at identifying qualified minority students than a race-blind model. The race-aware model picked up on the racial disparity in the underlying data (for instance, minority students underperforming in high school tests as compared to college tests) and compensated for it, enabling it to achieve better accuracy for both the overall population and the minority group.[23] This point was echoed by the UK's Information Commissioner's Office, which argued that disallowing AI models from using disability status "could mean the system is more likely to

22 Karen Hao, "Facebook's ad algorithms are still excluding women from seeing jobs" *MIT Technology Review* (9 April 2021) <https://www.technologyreview.com/2021/04/09/1022217/facebook-ad-algorithm-sex-discrimination/> (accessed 18 December 2021)

23 Jon Kleinberg *et al*, "Algorithmic Fairness" (2018) 108 *American Economic Association Papers and Proceedings* 22.

discriminate against people with a disability because it will not factor in the effect of their condition on other features used to make a prediction".[24]

23 It is clear from these studies that fairness through blindness does not work, especially against AI models which can pick up on the slightest correlations between the data and sensitive attributes. As organisations collect and use more data to train their AI models, the likelihood that at least one variable is correlated with the sensitive attribute also increases.

24 More worryingly, not collecting data on the sensitive attribute is also counterproductive to building fair AI models. Measuring how unfair a model is requires knowledge of how the model is discriminating based on the sensitive attribute, which in turn requires organisations to collect data on it. To illustrate the point, in the US, one measurement of bias that requires data on the sensitive attribute is the 80% test. In the Uniform Guidelines on Employee Selection Procedures, which is used to enforce the prohibition against discrimination in the Civil Rights Act of 1964, adverse or disparate impact is defined as a "substantially different rate of selection in hiring, promotion, or other employment decision which works to the disadvantage of members of a race, sex, or ethnic group".[25] The 80% test is a method adopted by the US courts to determine if there was a "substantially different rate of selection".[26]

25 Consider a hypothetical scenario of investigating Company A's hiring algorithm for bias against female applicants. The table below shows a breakdown of the key numbers: number of hires, number of applicants, and the hiring rate (calculated by dividing the number of hires by the number of applicants).

	No. hired	No. applied	Hiring rate
Men	150	300	50%
Women	60	200	30%

26 To apply the 80% test, one divides the hiring rate for women by the hiring rate for men. In this case, 30% divided by 50% is 0.6, or

24 *Guidance on the AI Auditing Framework: Draft Guidance for Consultation* (Information Commissioner's Office, 2020) at p 59. <https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf> (accessed 12 January 2022).

25 "Uniform Guidelines on Employee Selection Procedures" *U.S. Equal Employment and Opportunity Commission* <https://www.uniformguidelines.com/uniformguidelines.html#129> (accessed 29 January 2022).

26 See *Wards Cove Packing Co v Antonio*, 490 US 642 (S Ct, 1989) and *Ricci v DeStefano* 557 US 557 (S Ct, 2009).

60%. Since 60% is smaller than 80%, Company A's hiring algorithm has failed the 80% test and is considered to have an adverse impact on female applicants. In this situation, Company A may be investigated for potential discrimination.

27 It would have been impossible to assess whether the AI model was biased without the applicants' gender. This is exactly what happened with the Apple Card incident – Goldman Sachs did not use gender at all, which prevented it from identifying and correcting for biased credit scores.[27] In addition, it would be impossible to run simple counterfactuals[28] through the model without data on the sensitive attribute. An easy way to identify potential bias is to create synthetic profiles that differ only on the sensitive attribute, and test if the model predicts differently for these profiles.

28 In jurisdictions where it is not legally permissible to collect data on the sensitive attribute, some industry practitioners have attempted to use coarse-grained demographic information to measure potential algorithmic bias. However, this is very technically challenging and impossible in some cases.[29] The other option is to impute the missing data by using the same data to predict what the sensitive attribute was. For example, this is done by the US Consumer Financial Protection Bureau to assess compliance with fair lending laws.[30] However, these methods are not always accurate.[31] It is worth noting that the more effective the imputation is, the more likely it is that the AI model has also uncovered the same information from the data. In countries without such legal prohibitions, organisations may nevertheless remain hesitant to collect

27 Will Knight, "The Apple Card Didn't 'See' Gender – and That's the Problem" *Wired* (19 November 2019) <https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/> (accessed 12 April 2022).

28 Note that this differs from the idea of "counterfactual fairness" as espoused by Loftus *et al*, which draws on causal theory to work out fully fleshed out counterfactuals. See Loftus *et al*, "Causal Reasoning for Algorithmic Fairness" *arXiv preprint* (15 May 2018) <https://arxiv.org/pdf/1805.05859.pdf> (accessed 10 April 2022).

29 Holstein *et al*, "Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?" *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (4 May 2019) at p 8 <https://arxiv.org/pdf/1812.05239.pdf> (accessed 13 April 2022)

30 "Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity" *Consumer Financial Protection Bureau* (17 September 2014) <https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf> (accessed 1 February 2022).

31 James R Koren, "Feds use Rand formula to spot discrimination. The GOP calls it junk science" *The Los Angeles Times* (28 August 2016) <latimes.com/business/la-fi-rand-elliott-20160824-snap-story.html> (accessed 1 February 2022).

data on sensitive attributes, fearing that they may be exposed to liability for discriminatory outcomes.[32]

29 It is highly counterintuitive to argue that organisations must collect data on sensitive attributes in order to train AI models that do not discriminate. In particular, this runs contrary to the current shifts to requiring organisations not to collect such data for data privacy reasons. However, as argued in Part II, the reality is that humans and AI models operate differently, and to apply the same intuitions and rules to both is a mistaken endeavour. Providing AI algorithms with data on the sensitive attribute enables it to take into account differential treatment and correct for it where appropriate. On a practical level, collecting the data means that society can measure and penalise organisations that do not actively ensure they remain unbiased. As Peter Drucker famously commented, "if you can't measure it, you can't manage it".

B. Fairness has conflicting definitions

30 In 2016, ProPublica published an article on how a recidivism algorithm used in several US States, known as the Correctional Offender Management Profiling for Alternative Sanctions ("COMPAS"), was biased against black defendants. By obtaining COMPAS scores from 2013 to 2014 and matching them with public criminal records up till 2016, the ProPublica team could measure how well the COMPAS model predicted recidivism since it had actual data on whether the individual re-offended within a two-year period.[33]

31 The results made headlines, and for good reason. ProPublica found that "black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs 23 percent)", and that "white defendants who re-offended within the next two years were mistakenly labelled low risk almost twice as often as black re-offenders (48 percent vs 28 percent)".[34] In essence, ProPublica accused COMPAS

32 Bogen *et al*, "Awareness in Practice: Tensions in Access to Sensitive Attribute Data for Antidiscrimination" *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (27 January 2020) at p 497 <https://arxiv.org/pdf/1912.06171.pdf> (accessed 13 April 2022).

33 Julia Angwin *et al*, "Machine Bias" *ProPublica* (23 May 2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (accessed 10 December 2021).

34 Jeff Larson *et al*, "How We Analyzed the COMPAS Recidivism Algorithm" *ProPublica* (23 May 2016) <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> (accessed 10 December 2021).

of being harsher on black defendants who did not recidivate, and more lenient on white defendants who did re-offend.

32 Within two months, the company that developed COMPAS, Northpointe, released a research report rebutting ProPublica's main claims. In particular, Northpointe argued that ProPublica "focused on classification statistics that did not take into account the different base rates of recidivism for blacks and whites", and if it had done so, it would have found no evidence of racial bias.[35] Northpointe demonstrated that the COMPAS model fulfilled positive predictive parity, which is satisfied when the probability that an individual that was predicted to recidivate, but did not actually re-offend within two years, is similar for both white and black defendants.[36] Independently, a group of researchers from the Community Resources for Justice, a criminal justice think tank, corroborated Northpointe's findings using a different approach that focused on the model's overall accuracy for each race, and found that there was no statistically significant difference between the accuracy values by race. Both teams pointed out that the base rate of recidivism was higher for black defendants (51%) as compared to white defendants (39%), and as such "racial differences in failure rates across race describe the behaviour of defendants and the criminal justice system, not assessment bias".[37]

33 Despite using the same dataset, ProPublica and Northpointe drew opposite conclusions. The sharp public debate over COMPAS illustrated an important but hitherto overlooked point: there were different definitions of fairness which were incompatible with each other. Northpointe focused on equalising accuracy and sought to answer the question: "how many defendants were correctly classified as higher risk out of all defendants who were predicted to be higher risk?". In contrast, ProPublica prioritised equalising outcomes from misclassification, and was looking to answer the question: "how many defendants were

35 William Dieterich, Christina Mendoza & Tim Brennan, "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity" *Northpointe Inc* (8 July 2016) at p 1 <https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf> (accessed 11 December 2021).

36 William Dieterich, Christina Mendoza & Tim Brennan, "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity" *Northpointe Inc* (8 July 2016) at p 11 <https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf> (accessed 11 December 2021).

37 Anthony W Flores, Kristen Bechtel & Christopher T Lowenkamp, "False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.'" (2016) 80(2) *Federal Probation Journal* 38 at 41 <https://www.uscourts.gov/sites/default/files/80_2_6_0.pdf> (accessed 11 December 2021).

incorrectly classified as higher risk out of all defendants who were actually low risk?"

34 In the algorithmic fairness literature, the two definitions used by Northpointe and ProPublica have been formalised into two concepts: sufficiency (outcomes should be independent of the sensitive attribute, conditional on the model's predictions) and separation (predictions should be independent of the sensitive attribute, conditional on the actual outcomes).[38] To illustrate this, consider two individuals who belong to different sensitive groups. Sufficiency is fulfilled when both individuals have the same prediction and are equally likely to be predicted correctly, while separation is satisfied when both individuals have the same outcome and are equally likely to be predicted correctly. The only difference here is whether to focus on individuals with the same predictions or the same outcomes. While this distinction may seem minor, the implications are significant when a given outcome (*eg*, recidivism) is more common for a particular group (*eg*, black defendants). Indeed, it was quickly demonstrated that these contrasting definitions of fairness were mathematically incompatible with each other unless the base rates of recidivism were equal.[39] Others have pointed out that there is a plethora of different fairness definitions, each focusing on different aspects of classification and misclassification.[40]

35 Beyond the fact that fairness has different definitions, it has also been highlighted that fairness does not come for free. Using the same COMPAS data, Corbett-Davies *et al* demonstrated that enforcing fairness would result in releasing some individuals that did eventually recidivate. In contrast, prioritising public safety by minimising recidivism cases leads to the same significant disparities in misclassification as ProPublica had reported.[41]

36 There are two important insights to draw from the COMPAS debate. First, it is not sufficient to simply require AI models to be fair – we must also articulate what kind of fairness we expect. Given the major socio-economic implications of these systems on society,

38 Solon Barocas, Moritz Hardt & Arvind Naranayan, "Classification" *FairMLBook.Org* (2019) <https://fairmlbook.org/classification.html> (accessed 29 March 2022).

39 Alexandra Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments" (2017) 5(2) *Big Data* 153.

40 Arvind Naranayan, "21 Fairness Definitions and their Politics" *YouTube* (23 February 2018) <https://www.youtube.com/watch?v=jIXIuYdnyyk> (accessed 1 April 2022).

41 Sam Corbett-Davies *et al*, "Algorithmic Decision Making and the Cost of Fairness" *Proceedings of the 23rd Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining International Conference* (13 August 2017) <https://5harad.com/papers/fairness.pdf> (accessed 1 March 2022).

we cannot leave the task of defining fairness up to private companies which are unaccountable to the public. Second, algorithmic bias can be detected without access to the underlying data or AI model. ProPublica's investigative work was possible with just a dataset of defendants and their COMPAS scores. Although this approach is not perfect,[42] one practical benefit is being able to circumvent the need to compel companies to release their proprietary models for examination.

IV. Examining different jurisdictions

37 This Part explores the current legal and regulatory landscape for AI models across four jurisdictions: Singapore, the EU, the US and China.

A. Singapore's approach

(1) Collecting data on sensitive attributes

38 The main legislation regulating private organisations' collection of personal data is the Personal Data Protection Act 2012 ("PDPA"). There are two main requirements for the collection of personal data[43] generally, which also apply, by extension, to data on sensitive attributes. First, the individual must have given, or be deemed to have given, his or her consent to the collection of his or her personal data.[44] Second, an organisation may only collect personal data about an individual for purposes that a reasonable person would consider appropriate in the circumstances and that the individual has been informed of.[45]

39 In practice, organisations have been urged not to collect data on sensitive attributes. For instance, in the employment context, the Tripartite Alliance for Fair and Progressive Employment Practices ("TAFEP") released the Fair Recruitment and Selection Handbook, which provides that "information about **age, date of birth, gender, race, religion, marital status** and **family responsibilities** including **whether**

42 Definitions of fairness that depend on observational data may not always correctly identify bias. See Solon Barocas, Moritz Hardt & Arvind Naranayan, "Classification" *FairMLBook.Org* (2019) <https://fairmlbook.org/classification.html> (accessed 29 March 2022).

43 Personal data is defined in s 2 of the Personal Data Protection Act 2012 (2020 Rev Ed) to refer to "data, whether true or not, about an individual who can be identified (a) from that data; or (b) from that data and other information to which the organisation has or is likely to have access".

44 Personal Data Protection Act 2012 (2020 Rev Ed) s 13. There are some exceptions where organisations may collect and use personal data without consent, as listed in s 17 of the Personal Data Protection Act 2012 (2020 Rev Ed).

45 Personal Data Protection Act 2012 (2020 Rev Ed) s 18.

an applicant is pregnant or has children, and disability should not be asked for in an application form"[46] [emphasis in original]. Even if there are specific requirements to ask for this information, the employer should state the reasons which should be job-related.[47] In the financial sector, the Monetary Authority of Singapore ("MAS") released the Principles to Promote Fairness, Ethics, Accountability and Transparency in the Use of AI and Data Analytics in Singapore's Financial Sector, which similarly provides that the use of personal attributes as input factors for AI-driven decisions must be justified. The illustration given is that age can be an input factor for an AI-driven decision given the relationship between age and retirement.[48] Evidently, the present understanding is that only sensitive attributes that are relevant to the decision made can be justifiably used as inputs. As explained in Part III.A, this may cause difficulties in assessing algorithmic bias and in developing fair AI models.

(2) Addressing algorithmic fairness

40 Singapore has taken a collaborative approach to regulating AI, with the Personal Data Protection Commission ("PDPC") issuing several guidelines co-created with private sector organisations. For instance, the PDPC has released two editions of its Model Framework – the first edition was issued in January 2019[49] and the second edition in January 2020.[50] The Model Framework is a voluntary algorithm-agnostic, technology-agnostic and sector-agnostic tool to enable organisations which are deploying AI solutions at scale to do so responsibly.[51] The two main guiding principles identified in the Model Framework are ensuring that first, the AI decision making process is explainable, transparent and fair, and second, the AI solutions are human-centric.[52]

---

46 *Fair Recruitment & Selection Handbook* (Tripartite Alliance for Fair & Progressive Employment Practices, March 2018) at p 21.

47 *Fair Recruitment & Selection Handbook* (Tripartite Alliance for Fair & Progressive Employment Practices, March 2018) at p 21.

48 *Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector* (Monetary Authority of Singapore, 2019) at p 7.

49 *Model Artificial Intelligence Governance Framework* (Infocomm Media Development Authority & Personal Data Protection Commission Singapore, 1st Ed, 2019).

50 *Model Artificial Intelligence Governance Framework* (Infocomm Media Development Authority & Personal Data Protection Commission Singapore, 2nd Ed, 2020).

51 *Model Artificial Intelligence Governance Framework* (Infocomm Media Development Authority & Personal Data Protection Commission Singapore, 2nd Ed, 2020) at paras 1.1, 2.2 and 3.2.

52 *Model Artificial Intelligence Governance Framework* (Infocomm Media Development Authority & Personal Data Protection Commission Singapore, 2nd Ed, 2020) at para 2.7.

41 On fairness, the Model Framework recommends that organisations should ensure that algorithmic decisions do not create discriminatory or unjust impacts across different demographic lines.[53] However, as elucidated in Part III.B, what is considered a "discriminatory or unjust impact" can vary depending on which definition of fairness one picks. Further, the Model Framework recommends enhancing the transparency of algorithms found in AI models, including explaining how deployed AI models' algorithms function and documenting the repeatability of results produced by the AI model where the results are not explainable.[54] However, while transparency is key for investigating the social, ethical and fairness concerns behind using algorithms,[55] transparency is not an end goal but a gateway to discourse about fairness. It is worth mentioning that the Model Framework does deal with biases in other forms. For instance, a suggested risk management and internal control measure is for organisations to understand the ways in which datasets may be biased and address this in their safety measures and deployment strategies.[56]

42 Similarly, the specific guidance issued by the MAS on Principles to Promote Fairness, Ethics, Accountability and Transparency in the Use of AI and Data Analytics in Singapore's Financial Sector[57] also does not discuss the issue of formulating fair AI algorithms. However, the Veritas consortium, formed by the MAS and financial industry partners, released a white paper in February 2022 which focused on articulating a fairness assessment methodology.[58] It explicitly discussed different fairness definitions and their trade-offs, and suggested a framework for

---

53 *Model Artificial Intelligence Governance Framework* (Infocomm Media Development Authority & Personal Data Protection Commission Singapore, 2nd Ed, 2020) at p 64

54 *Model Artificial Intelligence Governance Framework* (Infocomm Media Development Authority & Personal Data Protection Commission Singapore, 2nd Ed, 2020) at paras 3.25–3.44.

55 Rebecca Heilweil, "New York City Couldn't Pry Open its Own Black Box Algorithms. So Now What?" *Vox* (18 December 2019) <https://www.vox.com/recode/2019/12/18/21026229/nyc-ai-algorithms-shadow-report> (accessed 13 March 2022).

56 *Model Artificial Intelligence Governance Framework* (Infocomm Media Development Authority & Personal Data Protection Commission Singapore, 2nd Ed, 2020) at p 24. See also p 38, which discusses minimising selection bias and measurement bias in datasets.

57 *Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector* (Monetary Authority of Singapore, 2019).

58 "MAS-led Industry Consortium Publishes Assessment Methodologies for Responsible Use of AI by Financial Institutions" *Monetary Authority of Singapore* (4 February 2022) <mas.gov.sg/news/media-releases/2022/mas-led-industry-consortium-publishes-assessment-methodologies-for-responsible-use-of-ai-by-financial-institutions> (accessed 19 August 2022).

companies to determine the appropriate fairness metric for their financial products.[59] This is a step in the right direction.

(3) Applying existing anti-discrimination laws

43 In 2021, the Government announced its plan to legislate a workplace anti-discrimination law that will enshrine existing non-binding fair employment guidelines, which were issued by the TAFEP as described above.[60] Pending the release and passing of the bill, it remains to be seen how and whether the new anti-discrimination law can be effectively applied to deal with algorithmic decision making.

B. The European Union's approach

(1) Collecting data on sensitive attributes

44 The General Data Protection Regulation ("GDPR"), which came into effect in May 2018, governs data protection in the EU. It is one of the toughest privacy laws as it imposes stringent obligations on organisations worldwide as long as they target or collect data related to individuals within the EU. Moreover, the fines imposed for violations are extremely high.[61]

45 Article 9(1) of the GDPR specifically prohibits the processing[62] of personal data that "[reveals] racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership", also known as special category data. This prohibition applies unless one of the ten conditions in Article 9 of the GDPR is satisfied, and there is a lawful basis under Article 6 for the processing of personal data more generally.

---

59 *Veritas Document 3A – FEAT Fairness Principles Assessment Methodology* (MAS, Accenture & Swiss Re, 2022) <https://www.mas.gov.sg/-/media/MAS-Media-Library/news/media-releases/2022/Veritas-Document-3A---FEAT-Fairness-Principles-Assessment-Methodology.pdf> (accessed 19 August 2022).

60 Michael Yong, "Highlights: PM Lee's National Day Rally 2021 speeches" *Channel News Asia* (29 August 2021) <https://www.channelnewsasia.com/singapore/national-day-rally-2021-live-ndr-lee-hsien-loong-highlights-2130556> (accessed 29 March 2022).

61 "What is GDPR, the EU's New Data Protection Law?" *GDPR.EU* <https://gdpr.eu/what-is-gdpr/> (accessed 30 March 2022)

62 Processing is defined in Art 4(2) of the GDPR to mean "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction".

Some of these conditions include express consent from the individual or when there are reasons of substantial public interest.[63]

46 Where an organisation wishes to use special category data to assess whether its AI system is discriminatory, the UK's Information Commissioner's Office[64] ("ICO") posited that the organisation can potentially rely on the substantial public interest condition in Art 9(2)(g) of the GDPR, assuming there is domestic legislation which allows for the use of such data to ensure equality of opportunity or treatment,[65] akin to the UK Data Protection Act 2018[66] ("UK DPA"). However, where an organisation wishes to use special category data as an input for all automated decisions to ensure that the AI system does not discriminate based on such categories (as suggested in Part III.A), the ICO does not state that it will be permissible under the GDPR read with the UK DPA. The ICO simply opined that this is prohibited under the GDPR unless the organisation has express consent from the individual or can meet one of the substantial public interest conditions laid out in the UK DPA. However, the equality of opportunity public interest condition can no longer be relied on because the data processing in question is carried out to make decisions about a specific individual.[67] The other substantial public interest conditions identified in the UK DPA are unlikely to apply either.[68]

---

63 For an explanation of each exception under Art 9(2) of the GDPR, see "What Are the Conditions for Processing?" *Information Commissioner's Office* <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/special-category-data/what-are-the-conditions-for-processing/> (accessed 30 March 2022).

64 The ICO is the UK's independent body set up to uphold information rights. Even though the UK is no longer part of the EU, the GDPR has been retained in UK domestic law as the UK GDPR. Hence, the ICO's analysis of the GDPR is helpful in understanding how the GDPR may be applied in other EU jurisdictions as well.

65 *Guidance on the AI Auditing Framework: Draft Guidance for Consultation* (Information Commissioner's Office, 2020) at pp 57–58 <https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf> (accessed 12 January 2022).

66 The UK Data Protection Act 2018 provides that equality of opportunity or treatment, as a substantial public interest, is met if the processing is necessary for the purposes of identifying or keeping under review the existence or absence of equality of opportunity or treatment between groups of people specified in relation to that category of personal data with a view to enabling such equality to be promoted or maintained. See UK Data Protection Act 2018 (c 12) Sched 1, Pt 2, s 8.

67 *Guidance on the AI Auditing Framework: Draft Guidance for Consultation* (Information Commissioner's Office, 2020) <https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf> (accessed 12 January 2022).

68 The substantial public interest conditions in the UK DPA include: (a) statutory and government purposes; (b) administration of justice and parliamentary purposes;

(cont'd on the next page)

47 The GDPR also specifically addresses automated decision making by providing that individuals have a right not to be subjected to a decision based solely on automated processing, which produces legal effects concerning him or her or similarly, significantly affects him or her.[69] However, this right is subject to a few exceptions, including if the data subject had expressly consented to an automated decision-making process.[70] This is a relatively trivial requirement today as many organisations can seek explicit consent from users for them to be subjected to AI-driven decision making.

(2) Addressing algorithmic fairness

48 In contrast to Singapore's balanced approach, the EU has taken a stricter rules-based approach to regulating AI. Beyond releasing guidelines for organisations to consider, the European Commission went further to release a draft AI regulation in April 2021.[71]

49 In gist, the draft AI regulation establishes a risk-based framework, which bans some uses of AI, heavily regulates high-risk uses, and lightly regulates less risky AI systems. High-risk applications include AI technology used in employment (*eg*, curriculum vitae-sorting software for recruitment procedures), essential private and public services (*eg*, credit scoring denying citizens opportunity to obtain a loan) and administration of justice and democratic processes.[72] This dovetails neatly with this article's focus on AI applications with potential for allocative harm. Obligations that these high-risk AI systems are subject to include having high-quality datasets to minimise risks and discriminatory outcomes, and appropriate human oversight measures to minimise risk.[73] However, there is no mention of algorithmic fairness in the regulations.

(c) equality of opportunity or treatment; and (d) racial and ethnic diversity at senior levels of organisations.

69 GDPR Art 22(1).

70 GDPR Art 22(2).

71 *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* (2021/0106 (COD), 21 April 2021).

72 "Regulatory Framework Proposal on Artificial Intelligence" *European Commission* <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai> (accessed 2 April 2022).

73 "Regulatory Framework Proposal on Artificial Intelligence" *European Commission* <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai> (accessed 2 April 2022).

50 Beyond the draft AI regulation, there has been sporadic discourse about algorithmic fairness.[74] In a paper written for the European Parliament titled "A Governance Framework for Algorithmic Accountability and Transparency", algorithmic fairness was identified as a guiding purpose for transparency and accountability.[75] However, none of the proposals relate specifically to algorithmic fairness, especially the difficulties on settling on a particular definition of fairness.

(3) Applying existing anti-discrimination laws

51 Present EU anti-discrimination laws are not well-suited to effectively regulate algorithmic discrimination.[76] In brief, European non-discrimination law generally targets two types of discrimination, namely direct and indirect discrimination.[77]

52 Direct discrimination occurs when an individual is treated less favourably than another is, has been or would be treated in a comparable situation because of membership in a protected class.[78] There is no

74 We have also considered the European Commission's *White Paper on Artificial Intelligence – A European Approach to Excellence and Trust* (2020), but there was nothing therein acknowledging the difficulty with conflicting definitions of fairness and about algorithmic fairness.

75 *A Governance Framework for Algorithmic Accountability and Transparency* (European Parliamentary Research Service, April 2019) at p 10.

76 See for instance Sandra Wachter *et al*, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI" (2021) 41 *Computer Law & Security Review* 1; and Philipp Hacker, "Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law" (2018) 55 *Common Market Law Review* 1. While there is an article in the EU Charter of Fundamental Rights prohibiting discrimination, this article does not discuss it as it generally only applies to public bodies of the EU and Member States but not private organisations.

77 Sandra Wachter *et al*, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI" (2021) 41 *Computer Law & Security Review* 1 at 15; Philipp Hacker, "Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law" (2018) 55 *Common Market Law Review* 1 at 9–12. This article does not go into detail on the multiple EU directives on non-discrimination. To list a few, there is the Racial Equality Directive (Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin), the Gender Equality Directive (Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast)) and the Employment Directive (Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment an occupation).

78 See Art 2(2)(a) of the Race Equality Directive.

need to prove any intention,[79] and there is a closed list of exceptions to the prohibition of direct discrimination.[80] In the context of AI-driven decision making, conventional AI algorithms do not allow data scientists to stipulate how to treat specific individuals. Unless membership in a protected class is intentionally and explicitly coded into the algorithm as a penalising factor, direct discrimination is unlikely to apply to AI systems.[81]

53 Indirect discrimination is when an apparently neutral provision, criterion or practice places individuals of one protected group at a particular disadvantage compared with other individuals, unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary.[82] To prove that a group has suffered particular disadvantage, the Court of Justice of the EU generally requires statistical evidence that a particularly large proportion of those negatively affected by a measure are from a protected group, and the effect of the rule is significantly more negative than those experienced by other individuals in a similar situation.[83] Whether an otherwise discriminatory measure can be justified is a context-specific and fact-dependent inquiry.[84] In the context of automated decisions, this second type of discrimination is more relevant as AI algorithms may disproportionately affect individuals from a particular group,[85] as corroborated by the examples in Part II.A.

54 There are two main issues with applying existing EU anti-discrimination laws to AI systems. First, indirect discrimination

79 Justyna Maliszewska-Nienartowicz, "Direct and Indirect Discrimination in European Union Law – How to Draw a Dividing Line?" (2014) 3(1) *International Journal of Social Sciences* 1 at 42.

80 Justyna Maliszewska-Nienartowicz, "Direct and Indirect Discrimination in European Union Law – How to Draw a Dividing Line?" (2014) 3(1) *International Journal of Social Sciences* 1 at 44–45.

81 Sandra Wachter *et al*, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI" (2021) 41 *Computer Law & Security Review* 1 at 9.

82 See Art 2(1)(b) of the Race Equality Directive.

83 *Handbook on European Non-Discrimination Law* (European Union Agency for Fundamental Rights and Council of Europe, 2018) at pp 56–58.

84 "Indirect Discrimination" *European Foundation for the Improvement of Living and Working Conditions* (22 February 2019) <https://www.eurofound.europa.eu/observatories/eurwork/industrial-relations-dictionary/indirect-discrimination> (accessed 10 April 2022).

85 Sandra Wachter *et al*, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI" (2021) 41 *Computer Law & Security Review* 1 at 44–45; Philipp Hacker, "Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law" (2018) 55 *Common Market Law Review* 1 at 10.

implicitly relies on a specific definition of fairness, that of statistical parity. However, the EU has not determined a specific threshold that will expose an organisation to potential liability, which creates uncertainty for both businesses and consumers alike. Statistical parity may not be the most appropriate definition here either, given that it is insensitive to disparities in misclassification outcomes (much like Northpointe's metric for COMPAS). Second, the requirement of a legitimate aim is often trivial for AI-driven decision making.[86] One can easily claim that the AI models were optimised for predictive accuracy, which brings tangible value to the organisation but also disproportionate effects on minority groups. These problems highlight the unique challenges that AI systems pose to the EU's current anti-discrimination laws.

C. The United States' approach

(1) Collecting data on sensitive attributes

55 There is data protection legislation regulating specific sectors in the US, but there is no overarching and comprehensive data protection law.[87] As such, there is no coherent approach to the collection of data on sensitive attributes. For instance, in the field of consumer credit and housing markets, the Equal Credit Opportunity Act ("ECOA") prohibits discrimination by creditors against credit applicants on the basis of certain factors including race, colour, religion, national origin, sex and marital status.[88] The ECOA's implementing regulations generally prohibit creditors from collecting demographic data related to the protected class status of credit applicants.[89] There are two exceptions to the prohibition. First, a creditor may request an applicant to designate a sex-related title, so long as the request is optional. Second, creditors are allowed to collect the race of credit applicants to monitor that their practices are not discriminatory.[90] In assessing potential discrimination, the Consumer

86 Philipp Hacker, "Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law" (2018) 55 *Common Market Law Review* 1 at 11 and 16–20.

87 *Data Protection Law: An Overview* (Congressional Research Service, 25 March 2019) at p 7 <https://sgp.fas.org/crs/misc/R45631.pdf> (accessed 1 April 2022); "America should borrow from Europe's data-privacy law" *The Economist* (7 April 2018) <https://www.economist.com/leaders/2018/04/05/america-should-borrow-from-europes-data-privacy-law> (accessed 1 April 2022).

88 "The Equal Credit Opportunity Act" *United States Department of Justice* (24 September 2021) <https://www.justice.gov/crt/equal-credit-opportunity-act-3> (accessed 1 April 2022).

89 Equal Credit Opportunity Act (Regulation B) 12 CFR § 1002.5(b).

90 Equal Credit Opportunity Act (Regulation B) 12 CFR § 1002.5(b). Winnie F Taylor, "Proving Racial Discrimination and Monitoring Fair Lending Compliance" (2011–2012) 31 Rev Banking & Fin L 199 at fn 15.

Financial Protection Bureau will also review documents related to models and algorithms.[91]

56 In contrast, in the field of home mortgages, pursuant to the Home Mortgage Disclosure Act,[92] mortgage lenders are allowed to collect data on applicants' race, ethnicity and sex. Federal Reserve analysts then study the collected data to determine whether there are any unlawful practices by mortgage lenders that violate the ECOA.[93]

(2) Addressing algorithmic fairness

57 In November 2018, New York City enacted Local Law 49 of 2018, which aims to provide transparency in government agencies' use of algorithms. Significantly, this is the first law in the US that targets algorithmic bias and establishes a taskforce to review the city's use of automated decision systems. However, the taskforce faced difficulty in obtaining access to AI systems used by the city, and was criticised for making overly generic recommendations.[94] In the specific field of employment, the New York City Council adopted a law in November 2021 which requires audits of algorithms used by employers in hiring or promotion processes.[95] Companies must conduct an audit assessing the algorithm for bias based on sex, race or ethnicity, and are obligated to notify job candidates if they have been assessed by an algorithm.[96]

58 At the federal level, in February 2022, the Algorithmic Accountability Bill was introduced in the US Senate and in the House of Representatives. The bill aims to ensure transparency in and oversight of software, algorithms and other automated systems used to make

---

91 Debo P Adegbile *et al*, "CFPB Announces Expansion of Unfairness Analysis to Address Discrimination" *Wilmer Hale* (24 March 2022) <https://www.wilmerhale.com/en/insights/client-alerts/20220323-cfpb-announces-expansion-of-unfairness-analysis-to-address-discrimination> (accessed 15 March 2022).

92 Home Mortgage Disclosure Act § 1003.4(a)(10)(i) and § 1003.4(b)(1).

93 Winnie F Taylor, "Proving Racial Discrimination and Monitoring Fair Lending Compliance" (2011–2012) 31 Rev Banking & Fin L 199 at 202.

94 Colin Lecher "NYC's Algorithm Task Force Was A 'Waste', Member Says" *Verge* (20 November 2019) <https://www.theverge.com/2019/11/20/20974379/nyc-algorithm-task-force-report-de-blasio> (accessed 15 March 2022); Kate Kayne "New York Just Set a 'Dangerous Precedent' on Algorithms, Experts Warn" *Bloomberg* (13 December 2019) <https://www.bloomberg.com/news/articles/2019-12-12/nyc-sets-dangerous-precedent-on-algorithms> (accessed 15 March 2022).

95 New York City Council Local Law 2021/144; Khari Johnson, "The Movement to Hold AI Accountable Gains More Steam" *Wired* (2 December 2021) <https://www.wired.com/story/movement-hold-ai-accountable-gains-steam/> (accessed 10 April 2022).

96 Ellen Glover, "NYC May Regulate Hiring Algorithms. Here's What That Means." *Built in NYC* (14 January 2021) <https://www.builtinnyc.com/2021/01/14/nyc-hiring-algorithm-bill-julia-stoyanovich> (accessed 10 April 2022).

significant decisions affecting individuals' lives. The bill also requires companies to conduct impact assessments for bias, transparency, privacy and other factors when using automated decision systems to make important decisions.[97]

59 Mandating an audit or impact assessment to detect potential algorithmic bias is a good first step. However, regulators must be cognisant of the different definitions of fairness and their respective limitations. In New York City's laws and the proposed Algorithmic Accountability Act, there is no mention of how fairness should be defined in the audit process.

(3) Applying existing anti-discrimination laws

60 Similar to EU anti-discrimination law, current US anti-discrimination laws are also ill-suited to effectively regulate algorithmic discrimination. In brief, under various US legislations,[98] there are two types of discrimination that are prohibited, namely disparate treatment and disparate impact. Disparate treatment occurs when a requirement or practice was chosen by the decision-maker because of, rather than in spite of, its adverse effects on relevant group members. Disparate impact occurs where a requirement or practice has a disproportionate adverse effect on members of a protected group and the defendant cannot show that the requirement or practice is adequately justified by business necessity.[99] This bears strong resemblance to the EU's anti-discrimination laws, with the addition of the 80% test which the US courts have recognised as a smoking gun for potential disparate impact cases. The same problems of the lack of consideration of other fairness definitions and the triviality of fulfilling the business necessity requirements apply.

97 Algorithmic Accountability Bill s 4(a)(11)(B).

98 The Equal Protection Clause of the US Constitution and all civil rights laws prohibit disparate treatment, while some civil right statutes prohibit disparate impact. See Jon Kleinberg *et al*, "Discrimination in the Age of Algorithms" (2018) 10 *Journal of Legal Analysis* 113 at 121. Further, the US Federal Trade Commission has made clear that three present statutes extend to cover biased algorithms even though they do not contain express language regulating AI, namely the Federal Trade Commission Act, the Fair Credit Reporting Act and the ECOA. See Elisa Jillson "Aiming for Truth, Fairness, and Equity in Your Company's Use of AI" *Federal Trade Commission* (19 April 2021) <https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai> (accessed 15 March 2022).

99 Jon Kleinberg *et al*, "Discrimination in the Age of Algorithms" (2018) 10 *Journal of Legal Analysis* 113 at 122–123.

D. China's approach

(1) Collecting data on sensitive attributes

61 The main legislation regulating data privacy is the Personal Information Protection Law[100] ("PIPL"). To process sensitive personal information, organisations must obtain individuals' express consent.[101] Such sensitive personal information includes biometrics, religious beliefs, specific identity, medical health status, financial accounts, the person's whereabouts and personal information relating to minors under 14 years old.[102] Further, sensitive personal information can only be processed if there is a specific purpose, it is necessary and where strict protective measures are taken.[103] The affected individuals must also generally be informed of the necessity of processing their sensitive personal information and the impact it has on their rights and interests.[104] Similar to the GDPR, the PIPL has extraterritorial reach.[105]

(2) Addressing algorithmic fairness

62 There is a clear recognition that algorithms are value-laden. This is expressly recognised in its Internet Information Service Algorithmic Recommendation Management Provisions (the "Provisions"), which came into force on 1 March 2022.[106] Article 1 makes clear that the Provisions were formulated in order to, *inter alia*, "carry forward the Socialist core value view" and "safeguard national security and the social and public interest". However, the Provisions only apply to the use of AI recommendation technology to provide Internet information services (*eg*, the dissemination of information),[107] rather than to AI algorithms

---

100 For an English translation, see "Personal Information Protection Law of the People's Republic of China" *The National People's Congress of the People's Republic of China* <http://en.npc.gov.cn.cdurl.cn/2021-12/29/c_694559.htm> (accessed 15 March 2022).

101 Personal Information Protection Law (PRC) Art 29.

102 Personal Information Protection Law (PRC) Art 28.

103 Personal Information Protection Law (PRC) Art 28.

104 Personal Information Protection Law (PRC) Art 30. "China's Personal Information Protection Law" *Clyde & Co* (11 October 2021) <https://www.clydeco.com/en/insights/2021/10/china-s-personal-information-protection-law> (accessed 15 March 2022).

105 Personal Information Protection Law (PRC) Art 3.

106 For an English translation of the Provisions, see "Translation: Internet Information Service Algorithmic Recommendation Management Provisions" *DigiChina* (10 January 2022) <https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/> (accessed 15 March 2022).

107 Brenda Goh, "China Swoops on Algorithm in Latest Tech Clampdown" *Reuters* (27 August 2021) <https://www.reuters.com/technology/china-issues-draft-

(cont'd on the next page)

generally. Specific concerns that the Provisions address include the spread of misinformation, lack of user autonomy, perceived economic harms of price discrimination and online addiction.[108]

63 To address these concerns, the Provisions stipulate that individuals should be given a choice to switch off algorithmic recommendation services and a right to an explanation of the impact of such services that have a significant impact on the rights and interests of users.[109] Moreover, government regulators will conduct algorithm security assessments and give suggestions to correct discovered problems.[110] Additionally, algorithmic recommendation service providers are obligated to "advance the use of algorithms in the direction of good". There is not much focus or attention on fairness more generally.

64 Further, under the PIPL, where organisations use individuals' personal information to conduct automated decision making, such organisations must guarantee transparency in the decision making and that the results are fair and just.[111] However, there is no further explanation of what definitions of "fairness" or "justice" are adopted.

(3) Applying existing anti-discrimination laws

65 Although China does not currently have a general anti-discrimination law, there are various legislations that prohibit discrimination in particular areas, such as the Law on the Protection of Rights and Interests of Women, the Law on the Protection of Persons with Disabilities and the Labour Law.[112] For example, the Labour Law stipulates that women should enjoy equal rights of employment as

guidelines-internet-recommendation-algorithms-2021-08-27/> (accessed 15 March 2022); Ananaya Agrawal, "China Creates New Rules to Control Algorithm Recommendation Services" *Jurist* (5 January 2022) <https://www.jurist.org/news/2022/01/china-creates-new-rules-to-control-algorithm-recommendation-services/> (accessed 15 March 2022).

108 Sapni G K & Mihir Mahajan, "Understanding China's Draft Algorithm Regulations" The Diplomat (16 September 2021) <https://thediplomat.com/2021/09/understanding-chinas-draft-algorithm-regulations/> (accessed 15 March 2022).

109 Personal Information Protection Law (PRC) Art 17.

110 Personal Information Protection Law (PRC) Art 28.

111 Personal Information Protection Law (PRC) Art 24. "China's Personal Information Protection Law" *Clyde & Co* (11 October 2021) <https://www.clydeco.com/en/insights/2021/10/china-s-personal-information-protection-law> (accessed 15 March 2022).

112 "Taking Stock of China's Anti-Discrimination Legislation" *Institute for Security & Development Policy* (September 2020) <https://isdp.eu/publication/taking-stock-of-chinas-anti-discrimination-legislation/> (accessed 15 March 2022).

men.[113] However, such laws have been criticised for having few and ineffective enforcement mechanisms, and for being vague about what constitutes discrimination and how to establish it.[114] These factors will similarly make it difficult to address and audit AI systems for potentially discriminatory impact.

V. The way forward

66 This Part draws on the earlier discussion on algorithmic fairness and relevant regulatory approaches to recommend three proposals for integrating algorithmic fairness into AI governance mechanisms.

A. Deconflicting privacy regulations from artificial intelligence governance

67 Privacy, transparency, fairness and explainability are values many hope to see in any AI-driven system. However, the algorithmic fairness literature has consistently demonstrated that there is an inevitable tension between some privacy regulations and achieving fairness in AI. As discussed in Part III.A, including the sensitive attributes in AI algorithms is critical in developing fairer AI models and enabling robust fairness audits.

68 The dominant focus today in privacy regulation is on data minimisation, premised on the idea that collecting less specific data about individuals is the optimal approach. Yet, it is unclear if this still holds up against the rise of big data. For one, removing direct identifiers is insufficient, as evidenced from a 2011 review of de-anonymisation attacks on healthcare data.[115] The US Census Bureau also ran a simulated re-identification attack on the 2010 Census records using additional data from data brokers, and were able to correctly re-identify 38% (or 52 million records), despite using more advanced disclosure avoidance

113 Vivian Wang, "China Moves to Overhaul Protections for Women's Rights, Sort Of" *The New York Times* (2 Jan 2022) <https://www.nytimes.com/2022/01/02/world/asia/china-womens-rights.html> (accessed 15 March 2022).

114 "China: Gender Discrimination in Hiring Persists" *Human Rights Watch* (29 April 2020) <https://www.hrw.org/news/2020/04/29/china-gender-discrimination-hiring-persists> (accessed 15 March 2022); "Employment Law Overview China 2019-2020" *L&E Global and Zhong Lun Law Firm* <https://knowledge.leglobal.org/wp-content/uploads/sites/2/LEGlobal-Employment-Law-Overview_China_2019-2020.pdf> (accessed 15 March 2022).

115 Khaled El Emam *et al*, "A Systematic Review of Re-Identification Attacks on Health Data" (2011) 6(12) PLOS ONE 1.

techniques like swapping.[116] For another, as discussed in Part III.A, plenty of proxy variables for sensitive attributes exist. Even if these proxy variables are only slightly correlated with the sensitive attribute, having enough of them enables a high-fidelity reconstruction of the sensitive attribute.[117] The cost of persisting with data minimisation as the primary privacy safeguard is that it hinders the development and auditing of fair AI models. Without data on the sensitive attribute, data scientists can neither assess how biased their models are nor work on improving the fairness of their models.

69 To clarify, this article is not calling for a dismantling of existing privacy regulations – in fact, they should continue to be strengthened in some domains. However, it is crucial to recognise that privacy and fairness do come into conflict in the era of big data and high performance computing. More in-depth discussions are required to determine how to best balance these values.

B. Articulating fairness definitions and standards for artificial intelligence governance

70 The COMPAS debate shows that it is insufficient to simply ask for AI models to be fair. Although this article focused on group fairness definitions such as statistical parity, predictive parity (or sufficiency), and equalised odds (or separation), there are also individual fairness and counterfactual fairness approaches that could not be covered in this article. Private actors are likely to pick what is convenient for their organisation, potentially exposing historically-disadvantaged groups to further harm. Moreover, putting humans in or over the loop does not necessarily lead to fairer outcomes, especially if they end up introducing their own biases into the system.[118] Accountability is no substitute for fairness.

71 How should society determine what definition of fairness to use? Selbst *et al* argue that "to treat fairness and justice as terms that have meaningful application to technology separate from a social context is … to make a category error".[119] Technological systems ultimately cannot be

116 Michael Hawes, "Understanding the 2020 Census Disclosure Avoidance System" *U.S. Census Bureau* (7 May 2021) <https://www2.census.gov/about/training-workshops/2021/2021-05-07-das-presentation.pdf> (accessed 15 February 2022)

117 Solon Barocas, Moritz Hardt & Arvind Naranayan, "Classification" *FairMLBook.Org* (2019) <https://fairmlbook.org/classification.html> (accessed 29 March 2022)

118 Ben Green & Yiling Chen, "Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments", paper presented at FAT* '19: Conference on Fairness, Accountability, and Transparency (29–31 January 2019).

119 Andrew D Selbst *et al*, "Fairness and Abstraction in Sociotechnical Systems" *Proceedings of the Conference on Fairness, Accountability, and Transparency*

(cont'd on the next page)

detached from the social, economic, historical and political contexts in which they are situated. What fairness means for automated hiring can be very different from AI-enabled marketing, particularly because the costs of misclassification are much more consequential in the former. Therefore, it may be worthwhile to begin by focusing attention on areas where significant allocative harm may occur, such as finance or recruiting. Having one definition covering more sensitive domains and another for the general usage of AI algorithms can be a good start. This echoes the EU's risk-based approach to regulating AI in its proposed Regulatory Framework, where high-risk applications of AI must meet more stringent requirements.[120]

72 After settling on a specific formulation of fairness, the next step is to develop industry benchmarks and audits for AI fairness. New York City has already taken the first step in mandating bias audits for automated employment decision tools using the 80% test as the passing benchmark.[121] Raji *et al* provide a thorough audit framework for AI, drawing on practices in other industries like aerospace, medicine, and finance.[122] Companies are likely to be interested in such audits as it sends a strong signal to investors and consumers that their products can be trusted. In fact, some companies have already started seeking external fairness and ethical audits for their AI algorithms.[123] However, without a clear and robust definition of fairness, these audits may not uncover the true extent of bias and may inadvertently become rubber stamps for flawed AI systems.[124] Defining fairness must precede auditing for fairness.

(January 2019) <https://dl.acm.org/doi/pdf/10.1145/3287560.3287598> (accessed 1 March 2022).

120 "Regulatory Framework Proposal on Artificial Intelligence" *European Commission* <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai> (accessed 10 April 2022).

121 "New York City Council Local Law 2021/144" *The New York City Council* (27 February 2020) <https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID> (accessed 14 April 2022)

122 Inioluwa Deborah Raji *et al*, "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing" *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (27 January 2020) <https://arxiv.org/pdf/2001.00973.pdf> (accessed 10 April 2022).

123 Jessi Hempel, "Want to Prove your Business is Fair? Audit Your Algorithm" *Wired* (9 May 2018) <https://www.wired.com/story/want-to-prove-your-business-is-fair-audit-your-algorithm/> (accessed 10 April 2022).

124 Hilke Schellmann, "Auditors are Testing Hiring Algorithms for Bias, But There's No Easy Fix" *MIT Technology Review* (11 February 2021) <https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/> (accessed 1 April 2022).

73 In the legal space, with a precise definition of fairness that can be mathematically operationalised in an algorithm, enforcement of anti-discrimination laws becomes a streamlined process. Instead of a lengthy trial to determine if an employer was discriminatory in the hiring process, courts can subpoena the AI algorithm and its related documentation and bring in experts to check how it was developed and whether biases were explicitly introduced into the model. This strengthens the justice system by ensuring that legitimate discrimination cases will be heard quickly and fairly, while also minimising the cost of potentially frivolous lawsuits.

C. Encouraging interdisciplinary sharing in artificial intelligence governance dialogues

74 There are many valuable insights from the field of algorithmic fairness that this article could not accommodate but that should also be discussed, such as removing all proxies of the sensitive attribute from datasets,[125] tackling bias in large language models,[126] and assessing whether fairness holds when AI systems interact and combine with each other.[127]

75 Apart from regular exchanges of ideas, this article proposes convening an interdisciplinary group comprising academics, practitioners, companies and regulatory officials from the legal, technical and policy fields to develop a framework for AI fairness for specific domain areas (as proposed in the earlier section). This group would define what algorithmic fairness should mean in these domain areas, what challenges we may have with operationalising it, and how these fit into the broader AI governance ecosystem. This is inspired by Singapore's Advisory Council on the Ethical Use of AI and Data, which similarly draws on experts and leaders from a variety of legal, business, and technical fields.[128] The Veritas initiative, led by MAS and comprising financial industry partners, is an excellent example of how this can work.[129]

---

125 Rich Zemel *et al*, "Learning Fair Representations" (2013) 28(3) PMLR 325.

126 Kellie Webster *et al*, "Measuring and Reducing Gendered Correlations in Pre-trained Models" *Google Research* (2 March 2021) <https://research.google/pubs/pub50755/> (accessed 3 April 2022).

127 Cynthia Dwork & Christine Ilvento, "Fairness Under Composition" *Fairness, Accountability, and Transparency in Machine Learning 2018* (15 July 2018) <https://arxiv.org/abs/1806.06122> (accessed 10 April 2022).

128 "Singapore's Approach to AI Governance" *Personal Data Protection Commission Singapore* <https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework> (accessed 15 April 2022).

129 "MAS Partners Financial Industry to Create Framework for Responsible Use of AI" *Monetary Authority of Singapore* (13 November 2019) <https://www.mas.gov.sg/

(cont'd on the next page)

76 Finally, algorithmic fairness can equip people with the right vocabulary to engage in meaningful discourse about fairness for AI systems. AI often comes across as a mythical black box to many. It is vital to demystify the inner workings of AI algorithms and what fairness means in an algorithmic context to enable more productive dialogue about fairness in AI. Beyond the technical aspects of algorithmic fairness, the AI governance community can also draw on research from the fields of sociology, philosophy, political science and economics to uncover the broader implications of AI as sociotechnical systems.

VI. Conclusion

77 Technology has long had different – and often adverse – impacts on minorities. Back in the mid-1950s, Kodak distributed a colour reference card known as the "Shirley" card, which came with a picture of a white woman to help photo labs calibrate skin tones and colours for printing. Unsurprisingly, this process performed poorly for people of other races, often under-exposing photos of darker-skinned subjects.[130]

78 Similarly, since AI algorithms learn from patterns in the data, AI models tend to use the majority as the reference point to the disadvantage of minority groups. Algorithmic fairness seeks to address this by reshaping AI algorithms to imbue fairness into its very core. In doing so, we discover the inevitable trade-offs between accuracy and fairness that we have to make.

79 This article has sought to bridge the gap between algorithmic fairness researchers and the AI governance community, so that promising research can be tested in practice and scaled up if successful. Technical approaches will never be the only answer, but they are often an integral part of the solution. As Sendhil Mullainathan mused, "it is much easier to fix a camera that does not register dark skin than to fix a photographer who fails to see dark-skinned people".[131] It was only four decades later, in

news/media-releases/2019/mas-partners-financial-industry-to-create-framework-for-responsible-use-of-ai#:~:text=The%20Monetary%20Authority%20of%20Singapore,and%20Data%20Analytics%20(AIDA).> (accessed 19 August 2022).

130 Mandalit del Barco, "How Kodak's Shirley Cards Set Photography's Skin-Tone Standard" *National Public Radio* (13 November 2014) <https://www.npr.org/2014/11/13/363517842/for-decades-kodak-s-shirley-cards-set-photography-s-skin-tone-standard> (accessed 11 April 2022).

131 Sendhil Mullainathan, "Biased Algorithms Are Easier to Fix Than Biased People" *The New York Times* (6 December 2019) <https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html> (accessed 10 April 2022).

1996, that Kodak introduced its first multiracial "Shirley" card. Let us not wait that long to start fixing our AI algorithms.