EDRi

EUROPEAN DIGITAL RIGHTS

Beyond Debiasing

—

Regulating AI and its inequalities

Report by Agathe Balayn and Seda Gürses, Delft University of Technology, the Netherlands.

We would like to thank **Seeta Peña Gangadharan** and **Wendy Grossman** for their extensive edits.

We also would like to thank **Piotr Sapieżyński**, **Donald Jay Bertulfo**, **Bogdan Kulynych**, **Reuben Binns**, **Martha Poon**, **Joris van Hoboken**, and **Lina Dencik** for their valuable comments. We would like to give a special nod to **Sarah Chander** and the rest of the EDRi members for their in-depth feedback on this report.

Index

Foreword

As the potential harms related to deployments of Artificial intelligence (AI) in all areas of public life are unveiled by critical researchers, investigators and civil society, there has been greater debate and awareness looking critically at the enhanced role of AI in our society, in particular its impact on marginalised communities.

EDRi has increasingly highlighted the harms AI systems pose to already discriminated people.[1] From predictive policing systems disproportionately assessing racialised peoples as presenting a higher risk of future criminality, various AI systems used to profile and extract data from people on the move, and the use of biometric mass surveillance practices for crime prevention likely to target marginalised communities.[2]

As a result, it has become necessary for policymakers to respond, advancing legislative and policy solutions to the discriminatory impact of AI systems. We realised, however, that the policy debate on AI and discrimination at EU level was increasingly structured around the concept of 'debiasing'. Whilst many of the concerns in this field raised by researchers and civil society, include, but are not limited to, flaws, errors and representational issues with espect to the composition of the datasets, the entire policy conversation centred on technical debiasing to the general exclusion of other governance solutions.

Underpinning this 'technocentric' approach is the assumption that the harms emanating from the introduction of AI can be primarily characterised as 'bias' or technical flaws in the system design. Following from this, all such problems can be addressed, mitigated and prevented using technical, 'debiasing' solutions, and it is the duty of the provider to do so.

With respect to structural discrimination, whilst some of the harms to marginalised groups may relate to biases in databases or in system design, the majority relate to how AI systems operate in a broader context of structural discrimination, recreating and amplifying existing patterns of discrimination. Framing the debate around technical responses will obscure the complexity of the impact of AI systems in a broader political economy and ringfence the potential responses to the technical sphere, centralising even more power with dominant technology companies. In the words of the authors, debiasing processes risk, "shifting political problems into the domain of design dominated by commercial actors".

The focus on 'debiasing AI' as the primary policy response to discriminatory AI may in fact serve to promote more uptake of AI systems that fundamentally discriminate, and worsen outcomes at individual, collective and societal levels. The authors of this report set out for us the boundaries and limits of what debiasing techniques in computer science can actually achieve, but also the broader, social, political and economic factors that technocentric approaches to AI and discrimination overlook. We are extremely grateful for their guidance, and hope this study will be useful to civil society and policymakers invested in structural responses to the harms AI can bring.

We should not allow techno-centric approaches to obfuscate more radical responses to the broad, structural harms emanating from AI systems. EDRi, along with a number of other civil society organisations, has called for governance, rather than technical responses to harmful AI systems. In particular, for 'impermissible' uses of AI that inherently violate fundamental rights, we have called for 'red lines' – prohibitions on such uses,[3] as well as/or a fundamental rights based approach to AI regulation, rooted in harm prevention and democratic oversight.

by Sarah Chander,
Senior Policy Adviser, EDRi

"This report was commissioned and reviewed by EDRi. However, it is not an EDRi position and does not necessarily reflect the stance of all EDRi members. The research was completed by Agathe Balayn and Seda Gürses of Delft University of Technology, the Netherlands."

Executive summary

AI-driven systems have broad social and economic impacts and demonstrably exacerbate structural discrimination and inequalities. For the most part, regulators have responded by narrowly focusing on the technocentric solution of debiasing algorithms and datasets. By doing so, regulators risk creating a bigger problem for both AI governance and democracy because this narrow approach squeezes complex socio-technical problems into the domain of design and thus into the hands of technology companies. By largely ignoring the costly production environments that machine learning requires, regulators condone an expansionist model of computational infrastructures (clouds, mobile phones, and sensor networks) driven by Big Tech. Effective solutions require bold regulations that target the root of power imbalances inherent to the pervasive deployment of AI-driven systems.

In summary, EU policy documents on AI show that to date policymakers have failed to genuinely engage with the structural discrimination brought by AI-based systems as well as the science of debiasing. Their technocentric approach empowers service providers as arbiters of discrimination and inequity, a paradoxical proposition. **Overall, current AI policy-making in the EU underestimates the inequalities that may materialise with AI and the way its application reinforces computational infrastructures in the hands of Big Tech.** In light of these shortcomings in AI policy-making, as well as other viewpoints presented above, we make six recommendations for policymakers, researchers, advocates and activists, and propose some broader frames for engaging technology companies going forward.

- **1. Policymakers adopting technocentric approaches to address the discriminatory impact of AI must define problems clearly, set criteria for solutions, develop guidance on known limitations, and support further interdisciplinary research.**
- **2. AI policies must limit the discretion of AI service providers in addressing discrimination and inequalities.**
- **3. AI regulation needs to go beyond ADMs, data and algorithms to include the spectrum of AI applications and the broader harms associated with the production and deployment of these systems.**
- **4. AI policies should empower individuals, communities and organisations to contest AI-based systems and to demand redress.**
- **5. AI regulation cannot be divorced from the power of Big Tech companies to control computational infrastructures. Addressing the rise of this infrastructural power requires long-term strategy and planning.**
- **6. AI regulation should protect, empower and hold accountable organisations and public institutions as they adopt AI-based systems.**

We base these recommendations on a three-part argument.The first part examines the current state of engagement of EU policymakers in questions of AI, discrimination and inequalities. We find that **key policy documents lack genuine engagement with existing theories, activism and laws around structural discrimination.** Policy documents erroneously use 'discrimination', 'equal access' and 'structural inequalities' interchangeably, and fail to ground these terms in existing EU law or social theory, or inform them with current social movements. The result is uncertainty in the scope of the problem to be addressed and the appropriateness of existing technocentric solutions.

EU policy documents favour debiasing datasets as the best means to address discrimination in AI, but fail to grasp the basics of debiasing approaches. When discussing debiasing, the documents mistakenly suggest that mitigating biases in datasets guarantees that future systems based on these so-called 'debiased datasets' will be non-discriminatory. Especially, proposed solutions focus on the establishment and monitoring of data requirements without providing much detail. They further fail to consider biases that may occur in models and their outputs, and show a lack of knowledge of existing relevant debiasing techniques and their limitations.

Whether applied to data-sets or algorithms, technocentric debiasing techniques have profound limitations: they address bias in mere statistical terms, instead of accounting for the varied and complex fairness requirements and needs of diverse system stakeholders. Regulators have neglected to grapple with the implications,

including technical impossibility results in debiasing (i.e., it is impossible to fulfill multiple debiasing requirements within the same model), and the flattening of differences, especially social and political differences that result from the pursuit of 'unbiased' datasets. For a given fairness objective, the performance of debiasing methods remains limited for statistical reasons, and the methods are hard to apply in practice given the need to collect sensitive data, and to understand the social context of a system.

Key EU policy documents appear to overstate the universal applicability of debiasing techniques, when in fact debiasing research addresses a limited, narrow set of social domains and a restricted set of machine learning techniques, often through a US-centric conception of discrimination and inequalities. The generalisability of the results to different domains, datasets and machine learning tasks is rarely studied and unlikely to hold.

Overall, policymakers do not provide sufficient guidance on debiasing requirements or how to address their techno - centric limitations. They also treat AI systems like a packaged product, pushing outside of regulatory scope the complexities of AI production pipelines and the continuously evolving services they deliver.

In sum, it is difficult to assess either the validity of the current policy focus on debiasing datasets or the future effectiveness of its application in regulating AI. In the second part, we show that even if policymakers develop a better grasp of the technical methods of debiasing data or algorithms, **debiasing approaches will not comprehensively or effectively address the discriminatory impact of AI systems. By design, debiasing approaches concentrate power in the hands of service providers, giving them (and not lawmakers) the discretion to decide what counts as discrimination, when it occurs and what it means to address it sufficiently.** Furthermore, when regulators rely upon debiasing as a solution to AI discrimination and inequalities, they divert our attention from the broader reordering of society, and inequalities, brought about by AI-based systems and the service providers who manage them.

Overall, given the limitations of debiasing techniques, policymakers should cease promoting debiasing as a silver bullet and instead advocate it only for the narrow applications for which it is suited.
In the third part, we combine our technical analysis with larger structural considerations to pinpoint the knock-on effects of hastily implemented and so-called debiasing-focused AI regulations. Specifically, we unpick the implicit and potentially problematic assumptions about phenomena inherent to machine learning and offer a sober assessment of the impossible tangle that debiasing-focused AI regulations create by absenting or abstracting away the production of AI. To do so, we develop and assess alternative viewpoints that go beyond current technocentric debates on data, algorithms and automated decision-making systems (ADMs). These viewpoints, which include machine learning, production,

infrastructural, and organisational approaches, have different implications for how comprehensively and effectively we assess AI, discrimination and inequalities.

For example, the machine learning view involves scrutinising the fundamental principles of machine learning applications. We survey the potentially harmful assumptions made when adopting machine learning more generally. These questionable assumptions concern reliance on and repetition of past data patterns, the soundness and desirability of the targeted task or inferences, the necessary discretisation of the environment into categories (i.e., creating discrete categories to capture elements of the environment in AI systems), and the problematic desire for universal scale. Precedents in machine learning applications, like the use of eugenics, phrenology and physiognomy in task modelling and the use of reductionist proxy attributes to represent complex categories like gender, race or sexuality, reflect implicit and socially unacceptable assumptions.

When applied to populations and social relations, machine learning applications can invoke the dark legacy of pseudosciences that are aligned with ideologies of social domination. We show that in machine learning, the design of the classification task may introduce representational and classification harms that may be reinforced by debiasing frameworks despite typically not being their focus. Even so, the machine learning view leaves open to question service providers' operational priorities, as well as the political and economic consequences of these systems.
In this sense, the machine learning approach to debiasing is powerful but remains a predominantly theoretical framework that is limited to assessing the way in which AI as a knowledge system may produce orderings and inequalities.

In presenting the production view, we enlarge the lens through which we examine machine learning's social, political and economic impacts. Here, we present machine learning as the output of digital production environments provided by the numerous actors in the supply chain of the business of computing.
In our assessment of scientific and regulatory literatures, we find that **neither scientists nor policymakers adequately grapple with the complexities of implementing technical solutions in the machine learning production pipeline.**
For instance, testing and mitigating bias only in training data risks missing other potentially harmful biases, such as when a model trained on this data is later fine-tuned and applied to different inference tasks.

The costly production environments needed for the development and use of machine learning are likely to increase individual and institutional dependencies on computational infrastructures. We draw attention to the role of Big Tech in assessing machine learning, discrimination and inequalities. Specifically, we argue that compute- heavy machine learning applications help to heighten societal dependency on computational infrastructure dominated by these companies.
If machine learning is implemented widely, it is likely to lead to greater concentration of technical, financial and political power in the hands of a few companies, inevitably raising global concerns around political, economic and social inequalities.

Finally, the organisational view of AI systems takes into consideration organisations' interdependence and path dependencies that arise from the entrenchment of computational infrastructures. Most AI-based systems will be deployed by and/or dependent on the computational capacities of Big Tech companies. In public sectors such as education, health, or transportation, integrating machine learning inserts computational infrastructures and their economic growth mandate into the heart of institutions tasked with serving the general public. Under these conditions, public institutions become both dependent on and instrumental to the economic success of technology companies: a co-dependency on unequal terms.

Beyond issues of financial dependency, the adoption of AI by public institutions cuts right into the execution of operations and the ability of these institutions to serve the public. **The impact of AI-based systems on the governance, operations and financial stability of public sector organisations is immense, and the integration of their everyday operations into current computational infrastructures could significantly transform, if not damage, the ability of public institutions to provide individuals with the necessary conditions in which to exercise their fundamental rights.**

Here are our recommendations in more detail (there is further discussion in Chapter E):

1. Policymakers adopting technocentric approaches to address the discriminatory impact of AI must define problems clearly, set criteria for solutions, develop guidance on known limitations, and support further interdisciplinary research.

1.1 Policymakers should engage with and learn from prior work on eliminating discrimination and inequalities as part of identifying and tackling AI's impact on inequalities.

1.2 Policymakers should better acquaint themselves with the basics and limitations of debiasing approaches before proposing them as solutions in regulatory interventions.

1.3 Policymakers should provide clearer guidance on applying debiasing and independent bias audits.

1.4 Policymakers should demand that any evaluation for discriminatory impact couples analysis of bias in an AI system's outcomes with an assessment of overall system objectives.

1.5 Policymakers should support interdisciplinary research on holistic approaches to auditing AI systems for discriminatory effects.

2. AI policies must limit the discretion of AI service providers in addressing discrimination and inequalities.

2.1 Policymakers should support an effective, decentralised system of assessing AI systems, discrimination and inequalities.

2.2 Policymakers should refocus the attention on bias and debiasing on to bias audits.

2.3 Policymakers should ensure that bias audits can be conducted independently.

2.4 Policymakers should set hard limits on access to sensitive data for bias auditing or debiasing.

2.5 Policymakers should avoid increasing surveillance of minorities or vulnerable populations in the name of debiasing or bias auditing.

3. AI regulation needs to go beyond ADMs, data and algorithms to include the spectrum of AI applications and the broader harms associated with the production and deployment of these systems.

3.1 Policymakers should expand the evidentiary scope of harms to non-technical criteria.

3.2 Policymakers should expand the scope of who (or what) may be classified as an affected party or AI subject and how they are harmed.

3.3 Policymakers should address distributed harms, exclusions and predatory inclusion through AI-based systems.

3.4 Policymakers should ensure that auditing extends across the supply chain of AI production and captures the evolution of AI-based services.

3.5 Policymakers should require that AI services available through application programming interfaces (APIs) are audited by service providers and organizations in each of the contexts in which they are deployed (and not only at the API).

3.6 Policymakers should bring harms accrued in the production of AI-based systems (e.g., labor conditions, environmental damate) into the scope of regulations.

3.7 Policymakers should ban the deployment of AI services that reproduce biological essentialisms and fascist, racist or supremacist conceptions of humans and societies.

4. AI policies should empower individuals, communities and organisations to contest AI-based systems and to demand redress.

4.1 Policymakers should enable the contestation and banning of harmful AI-based services.

4.2 Policymakers should enable affected parties to trigger internal and independent audits.

4.3 Policymakers should ensure that audits of AI systems include and empower affected parties.

5. AI regulation cannot be divorced from the power of Big Tech companies to control computational infrastructures. Addressing the rise of this infrastructural power requires long-term strategy and planning.

5.1 Policymakers should include within AI policy the broader impacts of the introduction of AI through computational infrastructures.

5.2 Policymakers should invest in research on the production of computational infrastructures and the political economy of Big Tech to capture AI's longer term impact on discrimination and inequalities.

6. AI regulation should protect, empower and hold accountable organisations and public institutions as they adopt AI-based systems.

6.1 Policymakers should grant rights of redress to organisations that deploy or are affected by third-party AI services and depend on computational infrastructures.

6.2 Policymakers should assess and build the capacity of public and private sector organisations to procure and deploy AI in a way that mitigates the technologies' broader harms and impact on inequalities.

---

1 **EDRi (2020). Recommendations for a Fundamental rights-based artificial intelligence regulation**
https://edri.org/wp-content/uploads/2020/06/AI_EDRiRecommendations.pdf

2 **EDRi (2020) Ban Biometric Mass Surveillance**
https://edri.org/wp-content/uploads/2020/05/Paper-Ban-Biometric-Mass-Surveillance.pdf

3 **EDRi (2021). Civil society call for AI red lines in European Union's artificial intelligence proposal**
https://edri.org/our-work/civil-society-call-for-ai-red-lines-in-the-european-unions-artificial-intelligence-proposal

If AI is the problem, is debiasing the solution?

Technological advances under the label of AI are seen by industry, governments and many other societal actors as cutting-edge and fundamental to innovation programs like digital transformation, the Fourth Industrial Revolution and the development of "smart" environments.

Whilst the development of AI has brought promises of increased efficiency - and even the ability of machines to simulate intelligence - it has also been widely accepted that AI can lead to harms.

The push to integrate AI into all domains of life, be it for employment, housing, education, health or policing, has raised concerns with respect to a diversity of harms, potentially at an unprecedented scale.

Those potential harms of AI pertaining to people have particularly caught the public interest. Across the globe, public institutions,[4] civil society groups,[5] researchers,[6] and social movements[7] have been reacting to reported potential harms due to the introduction of AI in digital services, and due to its organisational use in ADMs.

Experts have pointed out that these systems are often in violation of fundamental rights with regard to discrimination, privacy or stereotypical representations. These systems have also been shown to cause distributive injustices, widening economic inequalities.

Consequently, experts and organisations have warned that the introduction of AI-based services and ADMs may produce or amplify societal inequalities.[8] They have also cautioned that addressing this problem should de-centre technology, acknowledging that, "these systems connect to larger systems of institutionalised oppression".[9]

While policymakers in Europe have recognised the broad range of harms across different domains that accompany the introduction of AI and ADMs, their policy responses in framing the problem and the solution space have been comparably narrow.

In particular, policymakers and documents have primarily focused on the use of AI in ADMs and the potential discriminatory effects of such systems due to "bias in data" and "algorithms".

By looking at the problem through the lens of data and algorithms, policymakers have ended up focusing their attention on AI "products" that interface with citizens unfairly. This has overlooked the technical, institutional and economic arrangements necessary to bring AI into the world, and how these arrangements may cause inequalities.

What can be considered a technocentric view that pervades data and algorithms has manifested itself in policy documents and campaigns that

aspire to address matters of "(non-)discrimination" in Algorithmic or Automated Decision Making Systems.

In tandem, the solutions proposed in AI policy-making have often highlighted a new mechanism - "debiasing" - as a path forward. Discrimination and inequalities due to data collection and processing have been a concern since the 1960s and 70s. Policymakers have revisited the topic with every new digital product trend, whether "big data", "smart cities" or "digitalisation".

What has changed in policy-making around AI is the prominence of debiasing methods as the way to mitigate concerns around discrimination and inequalities that follow from the introduction of AI and ADMs.

To highlight the prominence of debiasing, it is sufficient to look at some of the recent policy documents coming out of European institutions. For example, the European Commission White Paper on AI states that:

> "[...] AI can ... lead to breaches of fundamental rights ... including non-discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation [...]. These risks might result from flaws in the overall design of AI systems (including as regards human oversight) or from the use of data without correcting possible bias (e.g. the system is trained using only or mainly data from men leading to suboptimal results in relation to women)."[10]

Similarly, the European Council Presidency's conclusions on The Charter of Fundamental Rights in the context of AI and digital change argues that AI may lead to less and more bias, stating that:

> "Data used to train AI systems therefore have to be accurate and adequate for their purpose and potential biases have to be addressed while allowing for sufficient flexibility in Research and Development for the further development of these systems.
>
> In this respect, we underline the importance of the principles of equality and non-discrimination in the design, development, deployment, use and evaluation of AI, particularly in systems integrating machine learning, and of ensuring that such systems are subject to adequate safe-guards and oversight, including market surveillance."[11]

---

4 **UN. 2020. New information technologies, racial equality, and non-discrimination**: Call for input. https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1593484

5 **EDRi. 2021. Civil society calls for AI red lines in the European Union's Artificial Intelligence proposal.** https://edri.org/our-work/civil-society-call-for-ai-red-lines-in-the-european-unions-artificial-intelligence-proposal **Hannah Couchman. 2019. Liberty's briefing on police use of live facial recognition technology.** https://www.libertyhumanrights.org.uk/wp-content/uploads/2020/02/LIBERTYS-BRIEFING-ON-FACIAL-RECOGNITION-November-2019-CURRENT.pdf **Privacy International. 2020. The SyRI case: a landmark ruling for benefits claimants around the world.** https://privacyinternational.org/news-analysis/3363/syri-case-landmark-ruling-benefits-claimants-around-world **AlgorithmWatch, Automating Society 2020 – Country issues Germany, France, Italy, Switzerland & Spain, Bertelsmann Stiftung, 2021.**

6 **Javier Sanchez-Monedero and Lina Dencik. 2020.** The politics of deceptive borders: 'biomarkers of deceit' and the case of iBorderCtrl. Information, Communication & Society (2020), 1–18. Available at: https://www.tandfonline.com/doi/full/10.1080/1369118X.2020.1792530

7 **Sadie Robinson. 2020. Furious students protest over A-Levels scandal.** https://socialistworker.co.uk/art/50488/Furious+students+protest+over+A+Levels+scandal. **Stop LAPD Spying Coalition. 2021. Stop LAPD Spying Coalition**. https://stoplapdspying.org **Isobel Asher Hamilton. 2019. Thousands of people across Europe are protesting and striking against Amazon on Black Friday.** https://www.businessinsider.com/amazon-strikes-and-protests-sweep-across-europe-on-black-friday-2019-11?r=US&IR=T.Yaseen Aslam and Jamie Woodcock. 2020. A History of Uber Organizing in the UK. http://oro.open.ac.uk/71933

8 **Virginia Eubanks. 2018.** The digital poorhouse. Harper's Magazine (2018).

9 **Seeta Peña Gangadharan and Jędrzej Niklas. 2019.** Decentering technology in discourse on discrimination. Information, Communication & Society 22, 7 (2019), 882–899. available at: https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1593484

10 **European Commission. 2020b.** White Paper on Artificial Intelligence: A European Approach to Excellence and Trust.

11 **Council of the EU. 2020a.** The charter of fundamental rights in the context of artificial intelligence and digital change.

The EU Commission's Anti-Racism Action Plan of September 2020 puts forward:

> "Specific requirements for the quality of training datasets and testing procedures for bias detection and correction that will serve to prevent negative discriminatory effects early on."[12]

In the recent EU Commission proposal for regulating AI, it is similarly mentioned:

> "Technical inaccuracies of AI systems intended for the remote biometric identification of natural persons can lead to biased results and entail discriminatory effects. This is particularly relevant when it comes to age, ethnicity, sex or disabilities."[13]

The terms used, such as "bias" and "debiasing", and the descriptions of the systems approaches to be taken, such as datasets, testing and tools, gesture towards "regulation by design" as a way to mitigate issues around discrimination and inequity associated with AI.[14]

For instance, the EU Commission says:

> "In order to protect the right of others from the discrimination that might result from the bias in AI systems, the providers should be able to process also special categories of personal data, as a matter of substantial public interest, in order to ensure the bias monitoring, detection and correction in relation to high-risk AI systems."[15]

In doing so, policymakers integrate technical concepts into policy documents. In the case of the use of the term "debiasing", policymakers do so as if they are proven solutions that can be standardised.

The emphasis in these documents on debiasing is also congruent with industry's marketing of debiasing (sometimes also called "fairness frameworks") as a necessary and sufficient protection of the discriminatory impact of the use of AI in ADMs.[16]

This document takes a deeper look at these policy proposals from the perspective of experts in computer science and systems engineering and asks the following questions:

- **What are the limits and harms of debiasing as a response to structural inequalities perpetuated through technology?**
- **Can debiasing be a helpful policy response to inequalities and discriminatory effects brought by the introduction of AI? If so, how and in which cases?**
- **What are alternative framings of the problem that could help better address potential inequalities and discriminatory impacts of AI?**

To set the scene, we first provide an overview of the ways in which important European policy documents frame the potential impact of AI on inequality, and the possible solutions to the resulting structural problems. We next provide an overview of debiasing approaches currently under research in computer science.

---

12 **European Commission.** 2020a. EU Anti-racism Action Plan 2020-2025. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-anti-racism-action-plan-2020-2025_en

13 **European Commission. 2021**. 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. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

14 **Karen Yeung, PE Vermaas, and Ibo van de Poel. 2015.** Design for the Value of Regulation. Handbook of Ethics, Values, and Technological Design (2015), 447–472.

15 **European Commission. 2021.** Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI act) and amending certain union legislative acts.

16 **Google. [n.d.]. Responsible AI practices.** https://ai.google/responsibilities/responsible-ai-practices, IBM. [n.d.]. IBM's multidisciplinary, multidimensional approach to AI ethics. https://www.ibm.com/artificial-intelligence/ethics; Microsoft. [n.d.]. Microsoft AI principles. https://www.ibm.com/artificial-intelligence/ethics.

Next, we move to deconstruct debiasing approaches as a solution to potential discriminatory harms of AI.[17]

We show the shortcomings of these propositions as well as limitations to their application. In later sections, we consider a number of alternative framings which may more comprehensively capture the impact of AI on societal inequalities, which is typically obfuscated by the attention on bias.[18]

These viewpoints allow us to demonstrate that AI and the increased dependencies on dominant computational infrastructures may intensify inequalities not only at the level of "algorithms", but more structurally through the reconfiguration of organisations, democratic institutions and economic relationships.

We conclude with recommendations to address the gap between current technocentric policy-making and approaches needed to grasp and address these broader harms of AI.

We hope that the document will help both policymakers and advocates to better capture the potential harms of AI, and support them in developing policies that ensure the application and deployment of these technologies is premised on their ability to bring greater economic prosperity, justice and equity to our future societies.

The structure of the report is as follows:

Part B: An explanation of what debiasing means in technical literature and in practice.

Part C: A description of the limitations of debiasing (and bias auditing) regarding the promises it makes of solving discrimination in the outputs of machine learning models.

Part D: A presentation of the discrimination and inequity-related harms that debiasing implicitly overlooks, in the shape of an alternative framing of AI inequity discussions.

Part E: A set of recommendations for policymakers to develop socio-technical solutions for addressing the harms that automated decision-making systems raise.

[17] **In line with our overall appeal to go beyond algorithmic or data-centric understanding of our AI condition, we have left a number of topics out of this report.** First, we intentionally divert from further approaches that locate the problems of AI only in the design of algorithms. In particular, we do not consider further algorithmic approaches like explainability and interpretability. While these frameworks have their differences, they have in common with debiasing approaches that they center the technology and service providers in the quest to improve the acceptability of AI based systems. We also deliberately leave out participatory design approaches to addressing the harms of AI. Proposals for participatory design, popular in industry and computer science, typically confirm a product-oriented approach, e.g., focusing on how we can produce more and better AI 'solutions' for and with users. While terms like 'refusal' have by now been integrated into the vocabulary of even big tech companies, participatory approaches are unlikely to foreground potential alternatives to developing or deploying these systems. Instead, we focus our energies on viewpoints that could help develop solidarities across communities, movements and political organisations in questioning the inequalities brought about by AI, our current computational infrastructures and the companies that power them.

[18] **Julia Powles. The Seductive Diversion of 'Solving' Bias in Artificial Intelligence.** https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

1. On our choice of terminology

Here, we clarify the terminology used in the rest of the report. **Artificial Intelligence (AI), Machine learning (machine learning), and Automated Decision Making systems (ADMs).**

As we explain in a latter section, policy documents generally talk about AI in a broad sense. While they often do not give a precise definition, we believe they refer to the broad set of computational methods that serve to perform a wide range of tasks automatically (e.g. detection and identification of objects in images, decisions on granting or rejecting loans requests, machine translation, etc.).

However, in computer science research, bias and debiasing are mainly discussed for systems relying on machine learning. Machine learning is a subset of AI methods that automatically "learn" and "improve" themselves based on data and/or simulated experiences. Because of this, we make the choice to discuss solely machine learning techniques in this report, instead of AI in general. We refer to AI only in reference to the way in which these technologies are discussed in policy-making.

Furthermore, most prominent examples around which debiasing research in computer science is organised focus on instances where machine learning is used for ADMs, such as systems for loan application acceptance or rejection, for recidivism prediction, etc.

> "Solely automated decision-making is the ability to make decisions by technological means without human involvement."[19]

In this report, we will discuss ADMs in the context of debiasing approaches, without necessarily assuming an entire absence of humans from the systems. In later sections, we will also extend our analysis to include the broader use of machine learning.

Fairness, bias, and discrimination

Most policy documents talk about issues of discrimination due to the application of AI systems and propose to solve these issues with debiasing methods.

In computer science, the convention is to refer to bias rather than discrimination:[20] datasets are biased, spreading biases in the outcomes of machine learning models - biases that might be harmful in discriminating ways when the models are applied to the real world.

While the terms can sometimes be found in literature, we will refrain from using the terms "unbiased data" or "debiased data" in the report, as we do not want to give the wrong impression that a dataset can be unbiased (it is always biased from one point of view or for some definition of bias).

Debiasing and bias auditing

Similarly, we believe that the terms "debiasing" and "auditing" can be misleading.

Debiasing refers to the application of select methods to address bias by achieving certain forms of statistical parity (e.g. making sure that the accuracy of a recidivism prediction system is similar for Black and White people by rebalancing a training dataset and re-training a machine learning model).

"Auditing", in computer science literature, in turn refers to evaluating whether these forms of statistical parity hold in a system. Yet, statistical parity does not necessarily bear any social understanding of discrimination.

(Un)Fairness

The computer science community has also been using the term (un)fairness to refer to these discriminating outcomes, in a rather interchangeable way with bias.

Conferences and workshops even use "fairness" in their name. In the report, we also make the choice to talk about bias and unfairness in interchangeable ways, in order to reflect more closely the computer science literature.

Structural discrimination

Whilst most institutional discrimination law focuses on concepts of unequal treatment on the basis of protected characteristics in individualised contexts, we use the concept of structural discrimination to refer to the broader societal conditions that generate individual instances of discrimination.

It is these structures that create and maintain vulnerability, harms and precarity aligned to constructed social, economic and political 'difference'. Structural discrimination is the intertwined relationship between historical injustices, epistemic (knowledge) erasure, laws, institutions, policies, practices, and social, political and economic disparities.

The effect of these factors is to further exclude and impose violence on marginalised people.[21]

Data, algorithm, and machine learning model

While algorithm and model are used interchangeably in policy documents, it is important to recall the difference. An algorithm is a process or set of rules to be followed to perform a calculation.

Machine learning *algorithms* are the set of calculations to perform in order to produce a machine learning model that will perform inferences regarding the future (e.g. predicting whether an individual is likely to recidivate).

These calculations are usually made on a set of training data: essentially, the machine learning algorithm identifies the main patterns in available data and guides the learning of an inference behaviour that copies and amplifies these patterns.

A machine learning *model* refers to the output of this process of algorithm execution. Concretely, it is a set of mathematical equations with parameters learned from the data using the algorithm, and which can now be used to make inferences on new data following the patterns learned from the training data. Following such definitions, data corresponds to the set of numerical information employed for executing the algorithm.

Model, system, pipeline

These words are also sometimes used interchangeably. Yet, while a machine learning model was explained above and is the main artefact that enables inferences to be made about any new data samples, additional components are also required to make such inferences.

For instance, these data samples need to be pre-processed to feed them to the model, and the outputs of the model might also require post-processing before presenting them to the model user. Hence, a *model* is a part of a larger *system*, that can make inferences.

Within the additional components that make up the system, we generally talk about a *data engineering* or *data processing* pipeline to refer to the chain of components that serves to process the data before making a new inference, or before training the model.

Prediction, inference, outcome, and output

When talking about the entities coming out of a machine learning model when presented with a data sample, the machine learning community often interchangeably uses the terms prediction, inference, outcome, and output.

We will also use these terms in the report. Yet, we emphasise the fact that such outputs are the results of a fixed set of calculations encoded into the machine learning model, these calculations themselves being the result of an optimisation algorithm that fitted to their labels as much training data as possible that it had previously been shown.

These outputs are not in any case the fruit of some random or illuminated guess/prediction on what the future will entail.

We also want to strengthen the conceptual difference between *outputs* and *outcomes*. The outputs of the systems are the inferences they make on new data. Yet, the systems are always used in an environment where these outputs will impact things or stakeholders belonging to this environment.

Thus, an outcome in this case refers to an output of a system and how it relates positively or negatively to a stakeholder.

Bias and debiasing frameworks always consider the outputs of the systems, however we believe (and we will show this in the rest of the report) that considering the outcomes is more relevant when accounting for potential discrimination caused by the systems.

---

19 **European Commission Data Protection Working Party (article 29). 2018.** Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation.

20 **We would like to warn the reader that the machine learning community also uses the term "bias" for talking about statistical concepts that have nothing to do with the societal biases at stake in this report** –more on this in subsection B.3.

21 **We have adapted a definition provided by Equinox Initiative for Racial Justice of 'structural racism' for these purposes: Equinox Initiative for Racial Justice.** "Towards Racial Justice: How the EU can create lasting change for racialised people" (2021). https://www.equinox-eu.com/wp-content/uploads/2021/03/WEB-EQUINOX-Towards-racial-justice-EU-institutions.pdf

B.

Current policy approaches to AI, discrimination and structure inequalities: A technocentric framing

1. Current EU discussions around bias in AI

Bias is a loosely defined term in policy documents, computer science research and practice. In this chapter, we first give an overview of the framing of bias in six policy documents published by European Union institutions.

We then give a technical account of how debiasing methods work in computer science research, and an overview of the contexts within which researchers discuss debiasing. The latter is intended to provide the reader with an overview of the approaches and techniques in computer science that are most likely intended by the calls for "debiasing" in European policy documents.

The interested reader can also find in Appendix C an explanation of the evolution of the terms "bias", "discrimination" and "fairness" as used in computer science.

From this chapter, the reader can get a sense of the contrast between the breadth of the bias problem as raised in policy documents, and the narrowness of its definitions and mitigation in computer science. The limitations of debiasing approaches also start to appear.

These are the foundations on which we explain the limitations of the bias framing in the following chapters.

Here, we provide an overview of the proposals made in the various policy documents published by EU institutions relating to AI and its societal harms, followed by a discussion of their limitations.

1.1. Sampling EU policy documents on AI

Most of the EU reports assume that using AI technology will primarily bring societal improvements.

They argue that these technologies can be useful in many domains: either to perform tasks more efficiently than humans do, such as in production systems or agriculture, or to perform tasks better (with higher accuracy) than humans, such as in healthcare or for climate change mitigation.[22]

They can also improve institutions, for instance by speeding up access to legal information, helping them take more objective decisions, or to assess fundamental rights compliance (with the assumption that accurate data lead to less biased decisions).[23]

They also discuss concerns and potential risks associated with the use of AI in real-life applications, as summarised in Table 1. The breadth of discussion about these issues varies per document.

While all mention discrimination (one of the main foci of our report), a few of them also mention adjacent problems regarding AI, including: its use in the context of justice; its dangers for privacy; equal access to AI for different populations; safety; and issues arising from the lack of transparency of the decisions taken by an AI system that therefore prevent full understanding and contestation.

In European law, discrimination, a complex social phenomenon, is captured using different terms. For instance, Fredman discusses that,

> "there are several different ways of conceptualising discrimination when it occurs on more than one ground.
>
> Terms such as 'multiple discrimination,' 'cumulative discrimination,' 'compound discrimination,' 'combined discrimination' and 'intersectional discrimination' are often used interchangeably although they might have subtly different meanings. There is no single settled terminology, either within legal systems or in the literature."[24]

Despite these many types of discrimination, the reports we considered often do not define "discrimination" (we report the exact words and explanations surrounding the use of the term "discrimination" in Table 1 for the reader to get an idea of the vagueness).

Table 1: Summary of the societal issues stemming from the application of artificial intelligence techniques, as mentioned in several EU documents.

	Discrimination	Justice	Privacy	Transparency/ explainability	Other
European commission white paper on AI *	Discriminating many people without the social control mechanisms that govern human behaviour	Right to an effective judicial remedy, a fair trial	Personal data protection, private life protection	Difficult to identify and prove possible breaches of laws	Freedom of expression, of assembly, human dignity, political freedom
EU charter of fundamental rights **	Perpatuate and amplify discrimination, including structural inequalities	-	-	Opacity	Fundamental rights, democracy, accessibility of services to citizens
EU anti-racism action plan ***	Perpetuate or stimulate racial bias, lead to biased results and ultimately to discrimination	-	-	-	-
Council of Europe - Preventing discrimination caused by the use of AI ******	Cause or exacerbate discrimination, denials of access to rights that disproportionately affect certain groups	Access to justice, fair trail, burden of proof, presumption of innocence	Right to private life and the protection of personal data	Discrimination difficult to prove, transparency and accountability regarding decisions	Equality of access to fundamental rights, to employment, education, housing, health, public service and welfare, [..] to digital tools
Data quality and Artificial Intelligence *******	Discriminate against individuals	Fair trial, effective remedies	-	Challenge a decision	Equal access to services
European commission proposal for a regulation on AI (AIA) ********	Lead to discrimination of persons or groups and perpetuate historical patterns of discrimination, or create new forms of discriminatory impacts	Right to an effective remedy and to a fair trial, right to defence and the presumption of innocence	Respect for private and family life, protection of personal data	Rights could be hampered, where such AI systems are not sufficiently transparent explainable and documented	Right to human dignity, freedom of expression and information, freedom of assembly, consumer protection, workers' rights, right to good administration, risk of harm to the health

European commission white paper on AI:** https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en. ****EU charter of fundamental rights:** https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights_en. ** EU anti-racism Action Plan:** https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-anti-racism-action-plan-2020-2025_en. ******Council of Europe - Preventing discrimination caused by the use of AI:** https://pace.coe.int/en/files/28715. ******* EU FRA - Data quality and Artificial intelligence:** https://fra.europa.eu/en/publication/2019/data-quality-and-artificial-intelligence-mitigating-bias-and-error-protect. ******** European commission proposal for a regulation on AI (AIA):** https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

Some reports mention equal access to services, some others mention structural inequalities, and others simply say "discrimination". They all mention lists of protected attributes which discrimination can apply to.

When further detailing these issues and explaining their causes – as summarised in Table 2 – all documents specifically highlight issues with the data on which the AI systems rely. They mention biases within this data and the extent to which the data represent different populations. They sometimes refer to the idea that data are biased due to reflecting existing societal biases (e.g. an unequal distribution of resources across groups of population), or due to biases of the humans who create the datasets (e.g. the human creators decide which data to include in or exclude from a dataset, impacting the breadth of populations reflected in the datasets).

A few of the documents, as demonstrated in the table, take a broader view and point at biases stemming from the algorithms that use this data, or at general issues with the overall process of developing AI systems from design to deployment. Mostly, the documents do not provide much detail, as shown in Table 2 which details the exact words used.

These policy documents also propose solutions or provide recommendations to tackle the issues they outline (these are summarised in Table 3). They often discuss the adaptation of existing legislation and the development of new regulations with different levels of precision.

Some documents also propose to create new policies to tackle structural issues around discrimination and diversity (diversity refers to both people and disciplinary input, e.g. computer science and social science) within education and industry, and to develop research on biases.'

While these are not elaborated further, all documents underscore the need for methods that ensure the production and use of 'less biased' datasets, and define data requirements that are to be monitored during development and possible deployment of AI systems.

Only a few documents also discuss the monitoring of whole systems. Even for the datasets, no in-depth guidance is provided on the ways to define these data requirements and evaluate them.

22 **European Commission. 2020b.** White Paper on Artificial Intelligence: A European Approach to Excellence and Trust.

23 **Council of the EU. 2020a.** The charter of fundamental rights in the context of artificial intelligence and digital change.

24 **Fredman, Sandra.** "Intersectional discrimination in EU gender equality and non-discrimination law." European Commission, DG for Justice and Consumers, Directorate D–Equality, Unit JUST/DI, Brussels (2016). https://op.europa.eu/en/publication-detail/-/publication/d73a9221-b7c3-40f6-8414-8a48a2157a2f

All documents underscore the need for methods that ensure the production and use of 'less biased' datasets, and define data requirements that are to be monitored during development and possible deployment of AI systems.

Table 2: Summary of the explanations provided in the EU documents about the causes for the discriminatory impact of AI discussed in Table 1.

	Flaws in the system design	Data	Algorithm
European commission white paper on AI *	Design, development, deployment	Bias, accurate and adequate for their purpose, data error, data fit for purpose	Biased algorithm
EU anti-racism action plan **	-	Data does not reflect the diversity of EU society	-
Preventing discrimination caused by the use of AI ***	Lack of diversity in companies	Data are by nature biased, choices about which data to use and which to ignore [...] as well as a lack of data on key issues, the use of procies and the difficulties inherent in quantifying abstract concepts	Optimized for efficiency, profitabily of other objectives without accounting for equality and non discrimination
EU FRA - data quality and artificial intelligence ******	-	Representation error: non-representative or biased data, structural differences in the data, measurement error, validity of target label representativity for task	-
European commission proposal for a regulation on AI (AIA) *******	Technical innacurancies of AI systems intended for the remote biometric identification of natural persons can lead to biased results and entail discriminatory effects	High data quality data sets should be sufficiently relevant, representative and free of errors and complete in view of the intended purpose of the system. They should also have the appropriate statistical properties, including as regards the persons or groups of persons on which the righ-risk AI system is intended to be used	Erroneous decisions or wrong or biased outputs generated by the AI system

European commission white paper on AI:** https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en. * EU anti-racism action plan:** https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-anti-racism-action-plan-2020-2025_en. *****Council of Europe - Preventing discrimination caused by the use of AI:** https://pace.coe.int/en/files/28715. ****** EU FRA - Data quality and Artificial intelligence:** https://fra.europa.eu/en/publication/2019/data-quality-and-artificial-intelligence-mitigating-bias-and-error-protect. ******* European commission proposal for a regulation on AI (AIA):** https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

Table 3: Summary of the propositions and recommendations mentioned by the EU documents to tackle the issues identified and listed in Table 1.

	Legislation	Debiasing data	Debiasing systems	Auditing data	Auditing systems	Transparency	Responsaibility	Other policies
European commission white paper on AI *	Adjust or clarify existing legislation	Training data: sufficiently broad, sufficiently representative	-	Verify training data	Human oversight; assessment repeated over time; the relevant programming and training methodologies, processes and techniques used to build, test and validate AI systems	Record keeping	Uncertainty as regards the allocation of responsabilites between different economic operators in the supply chain	-
EU charter of fundamental rights **	Legal and regulatory frameworks	No explicit mention of debiasing data	Specific requirements [..] for the design, development, deployment and use of AI	No specific guidance on data or system	Identifying, predicting potential impacts	-	-	Awareness about the use of techonologies, AI and legal literacy
EU anti-racism action plan ***	-	Requirements for the quality of training datasets, bias correction	-	Testing procedures for bias detection	Continuous monitoring [..] throughout the AI lifecycle	-	-	-
Council of europe - preventing discrimination caused by the use of AI ******	Ethical principles, regulations, international standards, review legislations	Unclear whether data or system	Non-discrimination in the design, procedures, tools, methods for regulating and auditing AI-based systems	No specific guidance on auditing data or system	Continuous "rigorous testing" before and after deployment	Transparency, including accessibility and explicability	Human responsaibility for decisions, including liability and the availability of remedies	Diversity in education and industry, digital literacy, interdisciplinary teams, debates
EU FRA - Data quality and Artificial intelligence *******	-	Unclear: probably implies that the data should be changed based on the assessment	-	Assessing data quality: question processes, relevance, representativity, coverage	-	-	Who is responsible for data collection, maintenance, dissemination?	-

European commission white paper on AI:** https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en. ****EU charter of fundamental rights:** https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights_en. ** EU anti-racism action plan:** https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-anti-racism-action-plan-2020-2025_en. ******Council of Europe - Preventing discrimination caused by the use of AI:** https://pace.coe.int/en/files/28715. ******* EU FRA - Data quality and Artificial intelligence:** https://fra.europa.eu/en/publication/2019/data-quality-and-artificial-intelligence-mitigating-bias-and-error-protect. ******** European commission proposal for a regulation on AI (AIA):** https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

1.2 Shortcomings of the EU documents with respect to debiasing

From these discussions, we identify and explain a number of shortcomings, contradictions and limitations in policy documents.

While discriminatory effects and inequalities are mentioned in the documents, the causes identified and the proposed recommendations only tackle a small subset of AI harms. That is, the focus on solely biased data and sometimes on human biases does not explain or include all forms of discrimination (e.g. structural discrimination) or other issues mentioned concerning justice or equal access to technologies.

"Correcting" for biases in the datasets will not detect, evaluate or solve many of these issues. Such shortcomings partially stem from the fact that the inequalities mentioned are not clearly defined and scoped in the documents.

Aside from the social complexities that cannot be addressed, debiasing methods do not ensure "fully" unbiased data due to technical limitations. Monitoring methods also come coupled with challenges pertaining to their technical limitations as well as their practical applications. These points are not accounted for in the policy documents.

The documents imply the existence of unbiased datasets or of an objective way to debias datasets. Yet, this is questionable as stakeholders do not all share the same vision of the desirable outputs to embed in a dataset. This complexity in defining the desired situation is obfuscated in the policy documents.

The focus on biases in datasets misses other types of biases that can arise from the choice of algorithms, optimisation metrics for these algorithms, evaluation metrics, etc.

The documents do not explain how to deal with the unavoidable trade-offs among the desired situations of different stakeholders, neither when debiasing datasets, nor when monitoring biases in systems.

Based on the points above, the documents appear to provide little guidance to developers for following the recommendations, are too vague for institutions to check whether they meet their contextual needs, and do not articulate clear requirements for policymakers and policy-enforcement to verify whether the recommendations are followed.

There are further complications when we drill down to the metrics, as we will return to in chapter C.

2. Problems that debiasing approaches aim to address

By leaning so heavily on debiasing, the above policy documents inadvertently conceive these approaches as a universal solution to the discriminatory impacts of AI. Such an erroneous conception is likely to lead to over-simplifications in the policy responses, neglecting certain harms emanating from AI systems.

Instead, it is important to first understand the conception of bias in computer science and the actual scope of debiasing, in order to identify and characterize the current mismatch between policies and computer science.

To expose the specific contexts in which debiasing research has been developed, we will answer: **(i)** for which type of applications do computer science researchers actually discuss bias and debiasing? **(ii)** what is considered bias in these domains? and **(iii)** what are the use cases typically scrutinised?

> **By leaning so heavily on debiasing, the above policy documents inadvertently conceive these approaches as a universal solution to the discriminatory impacts of AI. Such an erroneous conception is likely to lead to over-simplifications in the policy responses, neglecting certain harms emanating from AI systems.**

2.1 Bias in machine learning tasks

2.1.1 Automated decision-making tasks

In computer science, bias is often discussed in the context of systems performing automated decision-making tasks. These systems mostly rely on machine learning techniques and are typically evaluated on a small number of selected use-cases.

For computer scientists, automated decision-making tasks can be characterized as follows:

Classification: Classifying data samples into one or multiple classes (e.g. giving or rejecting a loan application).

Regression: Attributing a numerical value to each data sample (e.g. how likely someone is to commit a crime).

Other inference tasks: Other tasks recently explored, such as the translation of text from one language to another, or the automatic captioning of images.[25] These are examples of computer vision and natural language processing research.

Recommender systems: Biases are also discussed in systems that recommend entities to individuals (e.g. recommending drivers and passengers on ride-hailing apps);[26]

And in systems that select sets of entities or rank entities depending on search queries – information retrieval and data management research – (e.g. selecting a subset of candidates to interview for a job offer).[27] Such systems do not necessarily use machine learning.

Besides classifying the tasks based on their objectives or on the algorithmic techniques they require, they are also often differentiated based on the type of data they rely on. The data samples that the decision-making systems employ are generally tabular data[28], but they can also include image data, text data or videos. We give examples of these tasks in the next subsection.

2.1.2 Biases in machine learning tasks

In all of these machine learning tasks, bias is an issue that is either observed in the outputs of the developed systems (also termed the inferred labels of various data samples at deployment time) or in the internal representations the systems rely on (also termed the feature representation that is learned or adopted during model training, and used by the machine learning model to make inferences at deployment time).

For instance, a classification system can be considered biased if it mis-classifies certain population groups more than others (e.g. young people who would have repaid their loan get their loan application rejected more often than older people who also repaid them).

It could also be considered unfair if its outputs are different for any two similar individuals whose primary difference would only be a protected attribute (e.g. two similar individuals differing only in age should both either see their loan rejected or accepted for the system to be considered fair).

An image captioning system could be considered biased if the representations it learned stereotype certain populations, such as always associating male pronouns to images with computers or female pronouns to images with kitchens.

In this report, we focus primarily on biases related to the outputs of the systems as these are the ones studied more often in the bias and fairness literature.[29]

These biases are often attributed to biases in the data used to develop these systems. For example, the stereotypes are already contained implicitly in the data as it might be easier to scrape images of women in kitchens than men in kitchens when using the web for data collection.

The systems then automate and possibly reinforce these biases by learning over the biased data. Similarly, in set selection or item ranking tasks, biases are considered to arise in two ways.[30]

The data and the results of the systems reflect the world, which is considered "distorted" compared to the ideal situation (e.g. the scenario where young and old people would not repay their loan at the same rate so the system would give them loan at different rates, yet the ideal situation would be when the rates are equal), or the results of the systems differ from the original data, while this data is assumed to be representative of the ideal, "unbiased" world (e.g. the scenario where young and old people do repay their loan at the same rate and it is considered ideal to give them loans at the same rate, yet due to issues with the data, the system does not reflect this equal rate).

25 **Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019.** Mitigating Gender Bias in Natural Language Processing: Literature Review. Association for Computational Linguistics (ACL 2019) (2019).

26 **Sühr, Tom, et al.** "Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019. https://dl.acm.org/doi/10.1145/3292500.3330793

27 **Michael D Ekstrand, Robin Burke, and Fernando Diaz. 2019.** Fairness and discrimination in retrieval and recommendation. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 1403–1404.

28 Tabular data are typical data structured into rows for each individual they describe, and into columns for the different features that describe these individuals.

29 **Mehrabi, Ninareh, et al.** "A survey on bias and fairness in machine learning." ACM Computing Surveys (CSUR) 54.6 (2021): 1-35.

30 **Julia Stoyanovich, Bill Howe, and HV Jagadish. 2020.** Responsible data management. Proceedings of the VLDB Endowment 13, 12 (2020), 3474–3488.

2.2 Common use-cases in bias research

For each of these tasks, the scientific community employs typical use-cases for studying biases. We present an overview of these use-cases.

The use-cases may or may not reflect the ones policymakers imagine when creating policies suggesting debiasing as a solution. Especially for cases that have not been studied until now, more work is necessary in order to form policies that accurately tackle the problems identified.

2.2.1 Use-cases for machine learning tasks on tabular data

Such tasks span various fields of applications. We present three of them here, but more are discussed in the literature.[31]

Recidivism. Automated systems are employed in the US (and increasingly in European contexts) to infer the likelihood of recidivism among previously incarcerated individuals.[32] Judges use this information to make decisions on jail time or the granting of bail.

These systems rely either on regression tasks if a numerical risk score is outputted by the system (e.g. a rational number between 0 and 1, or an integer number between 1 and 10), or on classification tasks if a discrete set of labels is outputted (e.g. "will re-offend" and "will not re-offend").

They obtain historical data about past offenders and use this data to predict whether new offenders are likely to commit a crime again. The data typically consists of the defendant's demographic information (e.g. gender, race), background (e.g., presence of offenders in their family or friends), criminal history (e.g. number of prior offenses), administrative information about their case (e.g. case number, arrest date, zipcode), and whether they re-offended over a certain time period.

Angwin et al. from ProPublica have shown that COMPAS, the automated system developed by the company Northpointe, incorrectly attributes Black defendants a high-risk of recidivism far more often than it does for White defendants, and conversely it more often incorrectly attributes a low risk of recidivism to White defendants than to Black defendants.[33]

Northpointe considers its system unbiased because in total it makes similar percentages of errors for both Black and White defendants (a fairness notion called "accuracy equity"). These variations around unequal errors in the outputs of the systems are what is considered bias in machine learning systems. In machine learning literature, this type of problem is also coined as "unfairness".

Predictive policing. Automated systems are developed to determine where and by whom certain types of crime are likely to be committed. These systems are often used by police forces to decide how to allocate policing resources across cities in order to prevent crime.[34]

Place-based predictive policing systems use historical data about crimes in different neighbourhoods of a city, to teach an automated system to allocate police resources to patrol certain parts of these neighbourhoods.

New data are collected by the police in each neighbourhood they patrol and fed back to the system, which updates its algorithm to define where to allocate resources on the next day.

Bias here is technically represented as an unequal distribution of police forces across neighbourhoods (the policeare sent predominantly to certain areas compared to others).

While crime is also present in these other areas, their rates are underestimated as the police are not set to patrol these neighbourhoods and hence cannot report on these crimes. There is an issue of feedback loops in the automated system as it is fed data specifically on the patrolled neighbourhoods, and hence learns to police these ones to the exclusion of others.

Besides, as it relies originally on historical data, which are likely to reflect human and institutional discrimination relating to the locations where the police generally patrol, the systems are biased from the start and this bias is simply reinforced with use.[35]

Loan applications. Automated loan attribution systems predict the likelihood of an individual to default on a loan they applied for, in order to define who to give loans to.[36] This inference is made based on demographic information of various individuals and on their financial information, such as their credit history.

A biased system could be onethat predicts that White and Black defendants would default in dissimilar proportions, despite equal default rates inreality.

We will see later in the chapter that there exists a plethora of definitions of an unbiased system.

2.2.2 Classification tasks on image and text data

Facial recognition. Buolamwini and Gebru have shown that the outputs of multiple industrial gender classification systems based on face images are biased insofar as they are more likely to misclassify certain populations than others.[37] In this case they misclassify more often women than men, darker-skinned individuals than lighter-skinned individuals, and darker-skinned women in general.

Sentiment and hate classification. Biases in the outputs of systems classifying the sentiment of a sentence (e.g. positive or negative), and the classification of a text based on whether it is hateful or not, have been discussed in the natural language processing literature.

Particularly, non-toxic sentences referring to certain identities (e.g. "I am a gay man") are more often misclassified as toxic than sentences for other identities,[38] and sentences referring to people from certain (dominant) racial backgrounds and genders are always systematically attributed a more positive sentiment score.[39]

Machine translation, image captioning, text generation. Machine translation systems have also been shown to be gender-biased as they often translate female identity terms into male ones.[40]

Biases are also found in image captioning when a machine learning model systematically outputs an identity based on an incorrect observation (e.g. it predicts the presence of a woman in an image

because it identifies a kitchen in the image, instead of identifying the characteristics of the actual person in the image).[41]

Similarly, systems that are made to generate text automatically have been shown to exhibit gender, race and religion biases.[42] For instance, the GPT-3 algorithm is found to generate stereotypical sentences about Muslims and Islam, "reproducing and reinforcing an Orientalist vision" of the religion.[43]

The machine learning systems for performing such tasks rely on the encoding of the sentences into features, using what is called "word embeddings" - mappings from sentence words to numerical vectors. Research has discussed the biases contained in these embeddings, where bias is seen as stereotypical associations of certain word concepts and identity words, leading to harmful representations.

For instance, "mathematics" is closely associated to "man" and "arts" to "woman"; European-American names to "pleasant" and African-American names to "unpleasant";[44] "computer programmer" to "man" and "homemaker" to "woman".[45]

---

31 **Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020.** On the Applicability of machine learning Fairness Notions. (2020).

32 **Propublica. [n.d.]. How We Analyzed the COMPAS Recidivism Algorithm.** https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

33 **Ibid. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm**

34 **Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian.** 2018. Runaway feedback loops in predictive policing. In Conference on Fairness, Accountability and Transparency. Pmachine learningR, 160–171.

35 **Patrick Williams and Eric Kind. 2019.** Data-driven Policing: The hardwiring of discriminatory policing practices across Europe. (published by the European Network Against Racism) https://www.enar-eu.org/IMG/pdf/data-driven-profiling-web-final.pdf

36 **Sam Corbett-Davies and Sharad Goel. 2018.** The measure and mismeasure of fairness: A critical review of fair machine learning. (2018). https://arxiv.org/abs/1808.00023

37 **Joy Buolamwini and Timnit Gebru. 2018.** Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency. 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

38 **Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018.** Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 67–73.; **Ji Ho Park, Jamin Shin, and Pascale Fung. 2018.** Reducing Gender Bias in Abusive Language Detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2799–2804. https://aclanthology.org/S18-2005.pdf

39 **Svetlana Kiritchenko and Saif Mohammad. 2018.** Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. 43–53. https://aclanthology.org/S18-2005.pdf

40 **Marcelo OR Prates, Pedro H Avelar, and Lu s C Lamb. 2019.** Assessing gender bias in machine translation: a case study with google translate. Neural Computing and Applications (2019), 1–19. https://arxiv.org/abs/1809.02208

41 **Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018.** Women also snowboard: Overcoming bias in captioning models. In European Conference on Computer Vision. Springer, 793–811.; **Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017.** Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. https://arxiv.org/abs/1803.09797

42 **Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020).** Language models are few-shot learners. https://arxiv.org/abs/2005.14165

43 https://towardsdatascience.com/is-gpt-3-islamophobic-be13c2c6954f

44 **Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017.** Semantics derived automatically from language corpora contain human-like biases. Science 356, 6334 (2017), 183–186. https://arxiv.org/abs/1608.07187

45 **Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016.** Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems 29 (2016), 4349–4357. https://arxiv.org/abs/1607.06520

2.2.3 Recommender systems

Recommender systems are used in many contexts. For instance, within platforms for job recommendations where employers post job offers, job seekers are presented with a ranked list of jobs corresponding to their skills and expectations, outputted by such a system.[46]

There, researchers can talk about biases or unfairness for different stakeholders (termed "multi-sided fairness").

For instance, one fairness metric entails that job seekers who have similar skills should be recommended similar job offers disregarding their demographics, while similar job offers from different employers with different demographics should also be presented to the job seekers at similar ranks.

Other contexts where similar biases would be discussed are "sharing economy" recommender systems such as Uber and Airbnb, online advertising, online dating, etc. For instance, although not even using an algorithm, show a racial bias in acceptance of guests on Airbnb, with potential guests with an African-American name being 16% less likely to be accepted as a guest than English American names.[47]

In Italy, Deliveroo's recommender system has been shown to downgrade workers returning from a period of absence for any reason, consequently giving them access to fewer jobs in general, and assigning them to jobs with worse conditions.[48]

2.2.4 Sets and rankings

The rankings of information retrieved from search engines can also be biased. Kay et al. show, for instance, undesired biases in the Google image search functionality when querying occupation-related images, with systematic underrepresentation of women and stereotype exaggeration.[49]

Depending on the use and user of the ranking in the application, the number of ranked entities that are really accounted for varies, hence the ranked entities do not all receive the same exposure. Entities belonging to minorities, although ranking rather high, might then remain "hidden" from the users of the application.[50]

As for set selection, a typical example is again found within the hiring context. For instance, a list of potential candidates to a job is collected, and a system is tasked to filter this list to retain a smaller number of candidates.[51]

Bias could arise when only candidates from certain sub-populations are selected – what the data management community also terms "diversity" issues. It could also be when a diverse set of candidates is retained but these candidates might not all be the most suited within each of the sub-populations considered – what is termed "unfairness" in some publications.

---

46 **Robin Burke. 2017.** Multisided fairness for recommendation. (2017). https://arxiv.org/abs/1707.00093

47 **Edelman, Benjamin, Michael Luca, and Dan Svirsky.** "Racial discrimination in the sharing economy: Evidence from a field experiment." American Economic Journal: Applied Economics 9.2 (2017): 1-22. https://www.aeaweb.org/articles?id=10.1257/app.20160213

48 https://www.rivistailmulino.it/news/newsitem/index/Item/News:NEWS_ITEM:5480

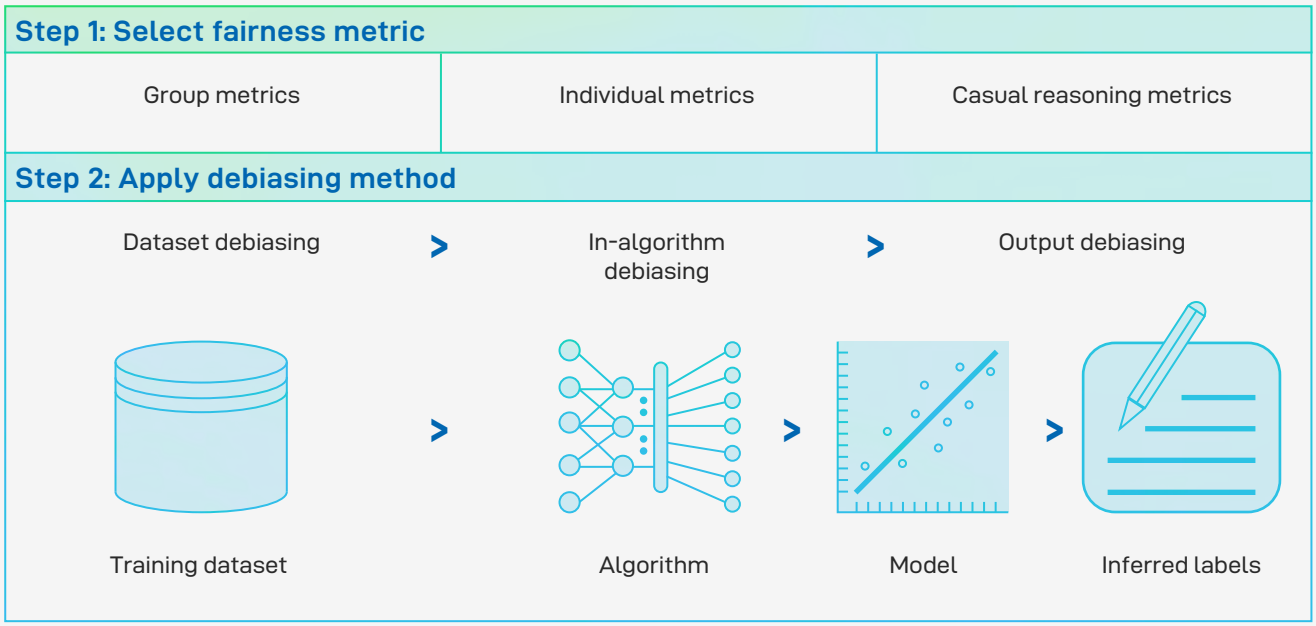
49 **Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015.** Unequal representation and gender stereotypes in image search results for occupations. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 3819–3828. https://dl.acm.org/doi/10.1145/2702123.2702520b

50 **Singh, Ashudeep, and Thorsten Joachims. "Fairness of exposure in rankings."** Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.; **Biega, Asia J., Krishna P. Gummadi, and Gerhard Weikum**. "Equity of attention: Amortizing individual fairness in rankings." The 41st international acm sigir conference on research & development in information retrieval. 2018.; Sapieżyński, Piotr, et al. "Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists." Companion Proceedings of The 2019 World Wide Web Conference. 2019. https://arxiv.org/abs/1802.07281

51 **Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. 2019.** Balanced Ranking with Diversity Constraints. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI. https://arxiv.org/abs/1906.01747

3. Employing debiasing workflows in practice

Figure 1: The debiasing workflow with the three types of bias conceptualisation and debiasing methods.



We now explain debiasing in more detail. We advise a reader who does not know about the basic machine learning setup and workflow to first read Appendix B (where we explain concepts such as confusion matrix, true positives and false negatives) in order to follow the concepts used in this section.

Social biases observed in the outputs of a machine learning model are typically mitigated using a two-step process.

First, a metric, usually termed a "fairness metric", needs to be selected. This metric, in theory, reflects the ideal outputs of the machine learning model when it is "unbiased". Then, a debiasing method is chosen according to the metric and applied to either the data, the machine learning algorithm or its outputs.

It transforms one of these in order to tend to fulfil the selected metric. This process is summarised in Figure 1. In the following, we explain in greater detail how these metrics and methods work.

3.1 Fairness metrics for sample / label biases

There exist various types of fairness metrics (or bias metrics), that can be divided into three main groups: statistical metric; individual similarity metrics; and causal metrics.[52]

3.1.1 Statistical group metrics

These metrics are qualified as group metrics because they rely on observing quantities related to the different inferences made by machine learning models for various groups of samples representing different groups of individuals.

Usually, someone who wants to evaluate a model's bias defines one or multiple protected attributes (also named sensitive attributes, these are variables for which it is considered relevant to monitor a model for bias, e.g. race and gender in facial recognition, age or marital status for loan applications, etc.) that are deemed relevant for the use-case at hand and that characterise the individuals on which inferences are made.

They then divide the available data into groups based on these attributes. Then, metrics are computed on these groups, as we detail below.

Error rates per group. These metrics consist of computing quantities related to a model's accuracy (e.g. numbers of correct or incorrect inferences for positive or negative labels, etc.) separately for two groups of population, and looking at their difference or ratio.

In order to compute a model's accuracy, practitioners need to assume the correctness of the labels in the available data, as measuring accuracy consists in verifying the extent to which a model's inferences match these expected labels (which would be meaningless if these labels were incorrect).

Example: Recidivism prediction - error rate metrics. The extent of bias over the race attribute may be gauged using the (positive) predictive parity measure.

This amounts to checking whether the occurrence of recidivism among individuals classified as high-risk is the same for both Black and White individuals. If Blacks are more likely to be incorrectly classified as high-risk than Whites, the prediction algorithm may be biased on the race dimension.

Another measure of bias is the error rate balance, which indicates whether an attribute such as race affects one's probability to be incorrectly classified as high-risk. For that, the ratios of false positives over the total number of "negative" individuals are computed for sets of individuals from different racial backgrounds.

Label distribution per group. Another set of statistical metrics is based on the likelihoods of obtaining positive or negative outcomes for different groups (termed demographic parity).

These metrics do not require any assumption on the available data labels as it only requires the inferences of the model. Compared to the above group of metrics, using these metrics reflects different notions of fairness.

Example: Loan attribution - Group metrics based on predicted label's likelihood.
In this example, the positive label is the acceptance of the loan application, and an unbiased model could be a model for which an equal proportion of women and men get such label.

One fairness metric would compute the likelihood of getting one's loan accepted (number of true positive and false negatives over the total number of individuals in the group) separately for men and women, and their ratio – disparate impact – or difference – statistical parity (these two metrics are two types of demographic parity).

Let us imagine that 200 men (and respectively 170 women) asked for a loan, and 100 of them (and respectively 70 of the women) had their loan accepted. The likelihood for men to get their loan application accepted is of 50% (100/200 = 0.5), and for women of 41% (70/170 = 0.41).

The disparate impact would then be 0.41/0.5 = 0.82 (the ideal disparate impact is 1), and the statistical parity 0.5 0.41 = 0.09 (the ideal statistical parity is 0), indicating the presence of bias.

3.1.2 Individual similarity metrics

The above statistical metrics rely on group measures, and do not account for relevant differences between the individuals of a same group. However, this may be important in many cases.

Example: Loan application - limitations of group metrics. A bank could respect statistical parity and give equal percentages of loans to its male and female clients, which would appear fair according to group metrics.
Yet, the men or women receiving the loans might not be the individuals who are most likely to repay them. The bank could also make decisions on men and women differently while still respecting statistical parity (e.g. choosing an adequate number of men at random, while choosing the women who are the most likely to repay).

That is why similarity metrics (i.e. individual fairness metrics) are proposed. In particular, causal discrimination is a metric that checks whether individuals that are identical, except for their protected attributes, receive the same labels.[53] Fairness through awareness is a relaxed version of causal discrimination and checks that similar individuals (for any similarity metric defined by the auditor) are treated similarly.

Example: Recidivism prediction - similarity metrics.
A model that gives a high-risk label to both men and women who committed x number of similar crimes would be considered unbiased by causal discrimination. A model that gives high-risk labels to all individuals who committed a similar number of similar offences would be considered unbiased by fairness through awareness.

---

52 **Sahil Verma and Julia Rubin. 2018.** Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). IEEE, 1–7. https://fairware.cs.umass.edu/papers/Verma.pdf

53 **Galhotra, Sainyam, Yuriy Brun, and Alexandra Meliou.** "Fairness testing: testing software for discrimination." Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering. 2017. https://people.cs.umass.edu/~brun/pubs/pubs/Galhotra17fse.pdf

3.1.3 Causal reasoning metrics

Figure 2: Example causal relations for the university admission scenario.

The above metrics rely only on the statistical relations between the protected attributes and the outputs of the models, missing out on the complexity of the causes for such relations. Instead, causal reasoning metrics model the causal relations between the attributes of a dataset, and observe fairness based on causal relations that one has to judge acceptable or unacceptable.

In practice, causal reasoning is still nearly impossible to use since it requires additional knowledge about the context of decisions and causal relations.

Example: University admissions - causal reasoning. Let us imagine a scenario where individuals apply to departments of universities, and an automated decision-making system decides which individuals to admit based on information about their qualifications. The causal relations can be modelled as follows in Figure 2.

Admissions are impacted by candidates' qualifications, but also by their gender (there could be discrimination based on gender in the data), and by the choice of department (some departments might receive more applications than others).

The choice of department might itself depend on gender (e.g. often the distribution of students' gender is not balanced within a department).

This causal model uncovers unfairness due to the direct relation between gender and admissions. If such relation did not exist, there could still be unfairness through the indirect relation between gender-department-admission depending on the context.

If candidates of one gender deliberately apply predominantly to departments with low rates of admission, it might be fair not to have an equal admission rate for candidates of different genders. However, if they apply to these departments due to peerpressure or historical stereotypes, for instance, then the existence of such causal relations should be considered unfair.

3.2 Debiasing methods for sample / label biases

Debiasing methods can be classified into three main groups, depending on which step of the machine learning pipeline they are applied to.[54] Debiasing can be done either by transforming the training dataset, by changing the method to optimise the machine learning algorithm, or by post-processing the outputs of the model in deployment. These are summarised in Figure 1.

The debiasing methods each improve on different fairness metrics, and account for different constraints of the usecases. For instance, someone who has access to the training data might prefer debiasing the data rather than the machine learning model to correct for biases earlier in their system. Someone who cannot modify the training process or data might prefer debiasing the outputs of the machine learning model. We present these different methods below.

3.2.1 Dataset debiasing

There are different methods for dataset debiasing, depending on the type of data that is inputted to the machine learning model and on the fairness metric selected.

Tabular data. A first idea is to remove any information about the protected attribute(s) from the dataset, i.e. simply removing the dataset features that correlate with the protected attribute. The assumption is that a machine learning model could not learn to treat different groups differently since the information about the groups would not be present.

Yet, this assumption has been shown to be flawed,[55] and instead can fail to prevent discriminatory outputs in deployed systems.[56]

Other attributes in the dataset might be correlated with the protected attribute(s) (these at-tributes are sometimes termed "proxy attributes") and challenging to identify, and hence the machine learning algorithm could still learn to infer behaviours that are indirectly based on the protected information.

For instance, "due to housing segregation, neighbourhood is a good proxy for race and can be used to redline candidates without reference to race. This is a relatively unsophisticated example, however. It is possible that some combination of musical tastes, stored "likes" on Facebook, and network of friends will reliably predict membership in protected classes."[57]

Besides, this approach poses difficulties, as the absence of information about the protected attributes acceptable makes it difficult to audit the models. This raises additional tensions because being able to employ debiasing requires access to these additional sensitive attributes which can be detrimental to certain groups – we discuss this in chapter C.3.2.

Other debiasing methods consist of transforming the training dataset to make it more similar to inferences that one establishes as "fair"/unbiased. This way, a machine learning model trained on such dataset should also learn to make fairer inferences.

The transformations of the dataset can be modifications of the values its features take, or of the weights attributed to each sample (for instance by resampling).

For instance, it is possible to artificially change the labels attributed to some of the samples (close to the decision boundary) in the dataset to increase statistical parity while keeping the loss in accuracy minimal.[58]

For example, if training samples of one group receive a lower rate of positive labels than the other groups, some of their labels can be switched from negative to positive to increase the rate of positive labels.

Extracting features from the dataset in a way that they retain the information contained in the data but also become independent of the protected attribute (called "fair" representations) has also been shown to improve on demographic parity and individual fairness measures.

In this case, the dataset itself is not modified but the features inputted to the algorithm are.[59]

> **Example: Loan application - dataset debiasing.**
> Let us imagine a dataset where 1500 women (and respectively 3000 men) applied for a loan and 300 of them (and respectively 900 men) were granted this loan.
>
> The disparate impact is hence equal to (300/1500)/(900/3000) = 0.67 (here women are less likely to receive loans than men). An "unbiased" dataset would show a disparate impact equal to 1, i.e. the same percentage of men and women would receive a loan.
>
> To get closer to such "unbiased" dataset, one can give more importance to data samples corresponding to the unprivileged population (here women) by repeating them in the dataset.
>
> For instance, by giving twice as much importance to some samples corresponding to women who are granted a loan – say to 100 of these samples – the disparate impact gets closer to 1: ((200+2x100)/1600)/(900/3000) = 0.83.
>
> Instead of repeating the samples, another option could be, for instance, to remove data from the privileged population, e.g. by removing 200 samples of men whoseloan got granted the disparate impact becomes (300/1500)/(700/2800) = 0.8.
>
> One more option could be to switch labels for some samples of men whose loan got accepted, or for some women samples whose loan got rejected. When the disparate impact approaches 1 for the training dataset, it is more likely that the model trained on it will also have a disparate impact nearing 1 (i.e. a "fair" model).

Non-tabular data. For non-tabular data, modified versions of the above methods apply. An approach similar to resampling is proposed.

Either more data are collected or created for the groups of population for which the machine learning model makes more mistakes – assuming that the more data there are, the more accurate the model will be - or data about certain groups are pruned to obtain lower accuracy in the inferences of certain groups in order to achieve accuracy rates closer to the one of the less privileged group.

The idea is to make the dataset more representative of the diversity of individuals on which the models make predictions, as did various companies that were publicly audited for race and gender biases in facial recognition,[60] or for non-binary genders.[61]

For biases in toxic sentence classification systems, datasets are resampled by artificially creating new sentences or by scraping additional sentences that balance the number of sentences with various identity terms and labels (e.g. toxic or not).[62]

However, this is limited since there can be an infinite number of identity terms, and practitioners would need to identify and mitigate all of them.

Transfer learning is also sometimes used. Here, a model is first trained on a large, "unbiased" dataset (which may not necessarily have been created for the target task). Then, it is trained on the available, biased dataset.

The idea is that, by training a model on an unbiased dataset, the problem of bias in the model will be addressed.[63]

3.2.2 Algorithm or output debiasing

Besides debiasing the dataset, other methods rely on modifying the machine learning algorithm and its training. Particularly, i) they incorporate the selected fairness metric into the objective function used to train the machine learning model;[64] ii) add constraints into the training process to account for the fairness metrics;[65] or iii) learn "fair" representations using adversarial learning methods.[66]

A last set of methods post-processes the inferences that the machine learning model makes, in order for them to approach closer the fairness metric that was selected. The exact methods vary per fairness metric. Kamiran et al. modify the inferences of the model that are the least sure in order to approach disparate impact,[67] while Lohia et al. propose a method for both disparate impact and individual metrics.[68] Hardt et al. change the labels based on certain computed probabilities to get closer to group fairness based on false negative and false positive rates.[69]

Example: Loan application - algorithm debiasing. Let us imagine that a machine learning model has been trained to output a number between 0 and 1, with numbers closer to 1 meaning that the individual is more likely to repay the loan, and numbers closer to 0 meaning that the individual is more likely not to repay the loan. 100 women have an output between 0 and 0.4, 100 women have an output between 0.4 and 0.6, and 100 have an output between 0.6 and 1 – respectively 100, 200, 200 for men.

If we consider that an output above 0.5 means a loan granted and otherwise a loan rejected, the disparate impact is equal to ((100 + 50)/300)/((200 + 100)/500) = 0.833. If however the samples with the most uncertain predictions were not necessarily given the label corresponding to the model outcome, but some of them got their labels shifted, a more ideal disparate impact could be reached.

For instance, a random fraction (let us say 50) of men with predictions between 0.4 and 0.6 can be shifted from positive to negative income, leading to a new disparate impact of ((100 + 50)/300)/((200 + 50)/500) = 1.0.

---

54 **Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019.** A survey on bias and fairness in machine learning. (2019). https://arxiv.org/abs/1908.09635

55 **Faisal Kamiran and Toon Calders. 2012.** Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems 33, 1 (2012), 1–33. https://link.springer.com/article/10.1007/s10115-011-0463-8

56 **Sapieżyński, Piotr, et al.** "Algorithms that "Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences." https://arxiv.org/abs/1912.07579

57 **Barocas, Solon, and Andrew D. Selbst.** "Big data's disparate impact." Calif. L. Rev. 104 (2016): 671. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899

58 **Faisal Kamiran and Toon Calders. 2012**. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems 33, 1 (2012), 1–33.

59 **Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013.** Learning fair representations. In International Conference on Machine Learning. 325–333.

60 **Inioluwa Deborah Raji and Joy Buolamwini. 2019.** Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 429–435. https://www.media.mit.edu/publications/actionable-auditing-investigating-the-impact-of-publicly-naming-biased-performance-results-of-commercial-ai-products

61 **Wenying Wu, Pavlos Protopapas, Zheng Yang, and Panagiotis Michalatos. 2020**. Gender Classification and Bias Mitigation in Facial Images. In 12th ACM Conference on Web Science. 106–114. https://arxiv.org/pdf/2007.06141.pdf

62 **Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018.** Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 67–73.; **Ji Ho Park, Jamin Shin, and Pascale Fung. 2018.** Reducing Gender Bias in Abusive Language Detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2799–2804.

3.3 Debiasing sample representations

We explained earlier that the inner representations (e.g. word embeddings) learned by machine learning models might also be considered biased.

Multiple approaches have been proposed to debias them, by removing the stereotypical information in existing embeddings,[70] or by training new embeddings with information related to the protected attributes constrained to certain vectorial spaces.[71]

3.4 Debiasing tools

Multiple debiasing toolkits have been built by several companies and academic research projects. The IBM AIF360,[72] the Microsoft FairLearn,[73] and the Aequitas framework of the University of Chicago,[74] all develop Python code foreasily applying various fairness metrics and debiasing methods, with varying degrees of guidance in the selection of metrics and in their understanding through visualisations.

The FairPrep framework is built on top of AIF360 to further facilitate its application.[75] Visualisation toolkits allow for the in-depth exploration of various metrics on various protected attributes and their combinations, like the Google What-If Tool,[76] FairVis,[77] and FairSight.[78] Google also proposes a fairness gym to simulate long-term fairness changes over various applications.[79]

63 **Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018.** Large scale crowdsourcing and characterization of twitter abusive behavior. In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 12.

64 **Bellamy, Rachel KE, et al.** "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias." https://arxiv.org/abs/1810.01943

65 **Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017**. A convex framework for fair regression. **Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017**. Fairness constraints: Mechanisms for fair classification. In Artificial Intelligence and Statistics. Pmachine learningR, 962–970. https://arxiv.org/abs/1706.02409

66 **Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. 2019.** One network adversarial fairness. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 2412–2420.; **Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019.** Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In Proceedings of the IEEE International Conference on Computer Vision. 5310–5319. https://ojs.aaai.org/index.php/AAAI/article/view/4085

67 **Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012.** Decision theory for discrimination-aware classification. In 2012 IEEE 12th International Conference on Data Mining. IEEE, 924–929. https://ieeexplore.ieee.org/document/6413831

68 **Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri.** 2019. Bias mitigation post-processing for individual and group fairness. In Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2847–2851.

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

70 **Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016.** Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems 29 (2016), 4349–4357. https://arxiv.org/abs/1607.06520

71 **Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang Chang. 2018.** Learning Gender-Neutral Word Embeddings. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. https://arxiv.org/abs/1809.01496

72 **Bellamy, Rachel KE, et al.** "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias." https://arxiv.org/abs/1810.01943

73 **Sarah Bird, Miro Dud k, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker.** [n.d.]. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report. Technical Report MSR-TR-2020-32, Microsoft, May 2020. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai

74 **Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018.** Aequitas: A **bias and fairness audit toolkit.** https://arxiv.org/abs/1811.05577

75 **Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. 2019.** FairPrep: Promoting Data to a First-Class Citizen in Studies on Fairness-Enhancing Interventions. https://arxiv.org/abs/1911.12587

76 **James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019.** The whatif tool: Interactive probing of machine learning models. IEEE transactions on visualization and computer graphics 26, 1 (2019), 56–65. https://ieeexplore.ieee.org/abstract/document/8807255

77 **Angel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019.** FairVis: Visual analytics for discovering intersectional bias in machine learning. In 2019 IEEE Conference on Visual Analytics Science and Technology (VAST). IEEE, 46–56. https://arxiv.org/abs/1904.05419

78 **Yongsu Ahn and Yu-Ru Lin. 2019.** Fairsight: Visual analytics for fairness in decision making. IEEE transactions on visualization and computer graphics 26, 1 (2019), 1086– 1095. https://arxiv.org/abs/1908.00176

79 **Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. 2020.** Fairness is not static: deeper understanding of long term fairness via simulation studies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 525–534. https://dl.acm.org/doi/10.1145/3351095.3372878

> **The idea is that, by training a model on an unbiased dataset, the problem of bias in the model will be addressed.**

C.

Deconstructing debiasing: A technocentric approach in the making

The previous chapter outlined the research efforts aiming to formalise the concept of bias in computer science, and to use debiasing tools to mitigate unfairness created from automated decision-making systems. The establishment of a community of researchers around the topic is already remarkable. Yet, despite much attention and financial support for the topic from Big Tech players, the field is still in its infancy.

A deeper look into debiasing research identifies a number of limitations in existing methods. Especially, their usecases are limited, the proposed conceptualisations of bias can oversimplify matters of discrimination, and the effectiveness and usability of debiasing methods and auditing tools are yet to be established.

Researchers and activists have further criticised debiasing for employing both a technocentric lens (as opposed to socio-technical systems or community-centric approach,[80] and a theoretical research lens (as opposed to a practical one[81]) on issues of discrimination in AI.

As a result, it is possible to argue that debiasing methods are not yet adapted to tackle discrimination in broader terms and in practice due to this current algorithm-centred view.

The limitations in debiasing methods contrast strongly with the public perception of the potential of debiasing in addressing discrimination and structural inequalities. The policy documents we sampled in Tables 1, 2 and 3 (chapter B) seem to place trust in debiasing approaches, suggesting that the immaturity of the field is not apparent to policymakers.

Their reliance on these approaches is also concerning since debiasing locates the problems and solutions in algorithmic inputs and outputs, shifting political problems into the domain of design dominated by commercial actors.

Table 4: Summary of the direct limitations of debiasing and bias auditing methods.

Simplifications necessary for the bias frame	Difficulties in applying methods in practice	Other limitations
Model-centric view Solely advocates for parity Trade-off between notions of parity that advantage different stakeholders Questionable definition of protected attributes Misalignment between system's outcome and human decisions **System-centric view** Limited impact of debiased system's outcomes on profound causes of discrimination Simplification of intersectional discrimination Fairness aspects unaccounted in debiasing Neglects negative externalities on the environment	**Limited performance of debiasing methods** Due to statistical nature of ML Due to dependencies in the ML pipeline **Practical challenges with the metrics** Anticipating potential harms Translating harms into relevant metric **Practical challenges with the use of data** Accessing information about the end-users Finding relevant data Raising harms when collecting data (e.g. privacy) Temporality of the required datasets and metrics	**Limited scope of debiasing methods** Limited to ADM Limited to a few applications **Dependence on service providers (SP)** Necessary incentives of the SP Complex responsability on multiple SP Centralisation of the value choices on SPs

In this chapter, we explore the theoretical and practical limitations of debiasing research, and the implications of these limitations for policymakers. We summarise our findings in Table 4 and below.

The scope of study of debiasing in computer science has been limited to automated decision-making systems (ADMs) that can directly impact human lives in a limited number of domains.

Yet, the discriminatory impact of AI is likely to apply more broadly, and to many more domains. We currently lack research that would reveal the potential and limitations of applying debiasing approaches to the inequalities that may arise in these other domains due to the introduction of AI.

This raises questions with respect to the plausibility of debiasing to mitigate discriminatory harms of AI as mentioned in policy documents.

Debiasing relies on conceptualisations of bias that do not capture the complexity of discrimination due to the limitations of the machine learning set-up. However, policy documents do not seem to be cognisant of such limitations.

Even if such conceptualisations are to mature over time, there are great challenges to the application of these conceptualisations in practice for auditing and debiasing purposes. These are not accounted for in policy documents. Current

debiasing proposals depend on service providers to implement solutions and audits. This raises serious concerns, as these providers may lack the incentives to address social inequalities.

Given the limitations of debiasing techniques, service providers may instead optimise debiasing outcomes to match their own interests. We have not found policy documents that discuss this allocation of responsibility and the alarming centralisation of value choices in the hands of service providers.

80 **Seeta Peña Gangadharan and Jędrzej Niklas. 2019.** Decentering technology in discourse on discrimination. Information, Communication & Society 22, 7 (2019), 882–899. https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1593484

81 **Dobbe, Roel IJ, Thomas Krendl Gilbert, and Yonatan Mintz.** "Hard Choices in Artificial Intelligence: Addressing Normative Uncertainty through Sociotechnical Commitments." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020. https://arxiv.org/abs/1911.09005

Current debiasing proposals depend on service providers to implement solutions and audits. This raises serious concerns, as these providers may lack the incentives to address social inequalities.

1. The scope of debiasing in computer science

1.1 The distinction between AI and ADM

We saw in the previous chapter that policy documents talk about debiasing in general terms to cover all matters of discrimination that may occur with the introduction of AI. This strongly differs from computer science research in two ways.

> **As an example, the game Pokemon Go was shown to place fewer Pokemons in rural areas and low-income neighbourhoods with racial minorities, creating a disparate "allocation of resources". This reflects assumptions about who has leisure time and how it is spent. This system leads to inequalities, not due to automated decisions applied to individuals, but due to the optimisation of the distribution of seemingly trivial digital objects.**

Firstly, policy documents do not refer to specific types of AI system when they discuss discrimination and bias (or at least they do not mention or use more specific typologies of AI techniques).

By contrast, most computer science research on debiasing specifically targets ADMs that rely on machine learning techniques to make decisions about individuals, or decisions that can impact individuals directly.

But, AI-based systems might create discriminatory harms due to a variety of applications, that do not fit in the mould of ADMs.

Machine learning can be used throughout digital services, for example to optimise the performance of a chatbot to improve efficiency, to test the colours on buttons to increase usability, or to recommend the film that will generate the greatest engagement from users, etc.

These are not examples of ADMs but still rely on AI and are only rarely discussed in debiasing literature. Overdorf et al. for instance mention that AI and inequality are not limited to employment, income, and housing allocations.[82]

As an example, the game Pokemon Go was shown to place fewer Pokemons in rural areas and low-income neighbourhoods with racial minorities, creating a disparate "allocation of resources". This reflects assumptions about who has leisure time and how it is spent. This system leads to inequalities, not due to automated decisions applied to individuals, but due to the optimisation of the distribution of seemingly trivial digital objects.

> **Example: Optimising user engagement - AI is greater than ADM.** When social networks use machine learning to optimiae their news feeds to generate greater engagement, they may end up down-ranking content by minorities, and up-ranking hate speech or misinformation.
>
> These social networks in reaction to criticisms or foreseen regulations on AI typically decide to use debiasing for toning down these criticisms. Debiasing approaches convert the issues into an issue of an automated decision system that decides whether to filter out posts from individuals with different political orientations, for instance, who produce misinformation or extremist posts.

The issue is then debiased by removing the same proportion of posts across user groups, or by having equal errors rates across groups. Yet, such a frame does not allow to be captured the way in which engagement optimisation leads to the erasure of minority voices or misinformation in the first place.[83]

The easy substitution of AI with ADMs and vice-versa can be misleading when it comes to the type of harms debiasing literature tackles.

For this reason, future policies would benefit from being more specific about the many ways in which the introduction of AI may lead to inequalities, and the limited ways in which debiasing can be used to potentially expose these effects.

1.2 The range of applications and domains studied in bias research

Policy documents tend to refer to debiasing as a catch-all method for any domain where applications of AI may be expected to have a discriminatory effect (see Table 2 Chapter B).

By contrast, regarding bias and debiasing, computer scientists refer to a very specific set of problems and techniques, as discussed in the previous chapter.

Such problems typically involve the allocation of resources in finance (e.g. loan application acceptance/rejection), justice (e.g. recidivism prediction for jail time/bail decisions) or hiring (e.g. selection of a candidate for a job), or the association of representational characteristics onto images or text (e.g. gender identification from facial images).

Due to this mismatch between the scopes of applications, we do not believe that debiasing can be easily adopted as a policy-response to any kind of discrimination for all AI systems.

In particular, despite the plethora of fairness metrics proposed up to now, it is not always the case that a metric exists for a specific machine learning task, and that a debiasing method has been developed for it.[84]

Certain issues do not receive as much interest as the issues that directly relate to individuals, such as tasks where protected attributes of individuals can be easily identified. For instance, conversational AI and image captioning both need a manual, tedious identification of what would serve as a protected attribute in sentences or images, instead of more automatic methods.

For example, one could define the association between gender (as apparent on the image – which is questionable) and various job-related captions as problematic (e.g. systematic association of images showing women to the label "nurse" or "housewife" and of men to the label "doctor" or "chef"), which would require the identification of both gender and the connected potentially problematic labels.

Even once an undesired bias is identified and a said-to-be ideal version of the model is imagined, debiasing methods cannot always be applied in practice since there do not necessarily exist methods for these issues.

We believe that researchers, policymakers and advocates should demand and source the study of AI applications in a wider variety of domains, such as ad budget allocation, online user engagement optimisation, and with extra attention given to the very differentiated ways in which discrimination and inequalities manifest themselves in context.

A more substantiated understanding of discrimination, including of how these dynamics may differ in Europe, can ensure the creation of appropriate policies and the proposition of relevant technical tools. This work cannot be performed by remaining solely in the computer science (CS) sphere, as CS researchers are not trained in understanding the societal contexts to model into their bias conceptualisations. We explain in more detail the dangers of limiting responses to a computer science view in the next chapters.

Until we have a better grasp of the impact of AI on inequalities in Europe, policymakers should be attentive to the limited scope of applications currently being studied and on which debiasing can be employed.

They should avoid recommending technical debiasing tools for problems on which these tools have not been tested.

> **Example: Computer vision applications – non-applicability of existing metrics and debiasing methods.** The Google computer vision API more often incorrectly associates hands of a darker skin colour holding a thermometer to guns, than lighter-skinned hands.
>
> This is not an easily foreseeable and measurable bias as it requires images of hands of various skin colours and holding various objects, checking how these objects are classified, and how offensive this is. Google could not improve its API, besides increasing the threshold for the label "gun" to be outputted.
>
> Another example where this API confused Black people with gorillas was "solved" by simply removing the label "gorilla" from the API, instead of improving the classification accuracy.
>
> Similarly, issues where different objects, that have various appearances (as they come from various countries) are not all recognised as accurately, would not be easily measurable as the various object appearances should be traced back to different cultures or countries, or any relevant protected attribute.[85]

2. Simplistic conceptualisations of bias

Debiasing methods aim at making the outputs of a system "fair", "unbiased", "non-discriminative". With technical definitions, it means that individuals who are similar based on protected characteristics should be treated similarly by the system, i.e. should receive the same outputs.

Yet, having different outputs is not necessarily what makes discrimination. Instead, it is often more the way these outputs impact differently upon the different individuals (potentially of the same protected characteristic) in the environment.

Debiasing relies on conceptualisations that cannot capture the complexity of discrimination due to the limitations of the machine learning set-up. Computer science researchers develop methods centred solely around the inputs and outputs of the machine learning models. However, when the system is used in an actual environment, its outputs might be used differently by different stakeholders, and the actual outcomes of the system might be different for different elements of the environment.

Focusing on outputs instead of outcomes cannot then accurately reflect the discrimination issues that take place. Impossibility theorems even show that various fairness conceptualisations are mutually exclusive, in which case stakeholders with different fairness values cannot all be satisfied.[86]

For instance, in the loan application example, the bank and its clients might not all agree on what fair decisions are, hence they might regard as important different fairness metrics.

Yet, the impossibility theorems show that only the metric aligned with the bank values or with a part of its clients' ones can hold at the same time.

Such simplifications are necessary when taking a technocentric approach to the problem of discrimination in ADMs in order to allow for the operationalisation of bias. But they leave out the real social context of the systems and, in certain cases, might reinforce harms more than address them (see below for further explanations).

Due to these simplifications, there is a gap between discrimination as understood by policymakers, and what discrimination means following the existing machine learning setup. This gap is not considered in policy documents that propose to resolve discrimination through debiasing.

In this section, we strive to provide examples of this gap. We start from the limited set of design choices available to a machine learning practitioner to formalise "discrimination", and we compare the resulting conceptualisations to a systemic view of the machine learning models in their environments of application.

The section intends to provide an explanation for why the claim that a system can be "unbiased", or "fair", according to a single metric is a far cry from a system free of discrimination.

2.1 Model-centric view of discrimination

We expect that the focus on machine learning models' outputs in fairness metrics is due to pragmatism. Computing fairness metrics typically requires accessing the outputs of the models (and possibly the ground truth information about data samples) and the sensitive attributes associated to each data sample. In other words, it only requires accessing the smallest set of information that is almost readily available to practitioners.

The metrics rely on checking some simple notions of parity between aggregates on this information (e.g. equal rates of getting a positive output across two groups of population corresponding to two sensitive groups), which does not require any additional contextual information.

By contrast, if one tries to check for a certain inequality between aggregates (e.g. the rates of getting a positive output should be twice as high for one group than the other), they would first need to establish a value for this inequality by translating contextual information into a meaningful and mathematically relevant value - which can be a challenging task to perform.

While such metrics are practical, they do not reflect the different stakeholders' desired conceptualisations of fairness. That is what we explain further in the next subsections.

82 **Rebekah Overdorf, Bogdan Kulynych, Ero Balsa, Carmela Troncoso, and Seda Gürses. 2018.** Questioning the assumptions behind fairness solutions.

83 https://arxiv.org/abs/1811.11293

84 **Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019.** Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–16. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation

85 **Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019.** Does object recognition work for everyone?. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 52–59. https://arxiv.org/abs/1906.02659

86 **Alexandra Chouldechova. 2017.** Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data 5, 2 (2017), 153–163.; Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Schloss Dagstuhl- Leibniz-Zentrum fuer Informatik. https://arxiv.org/abs/1610.07524

2.1.1 Parity as the unconditional desired outcome

Firstly, the metrics fail to account for applications where parity (technically defined as equality of outputs for similar individuals or groups of individuals) is not necessarily wanted for certain stakeholders.[87]

> **Example: Loan application - Limitations of parity.** In our loan application example, achieving disparate impact might lead disadvantaged minorities to obtain loans that they are unable to repay, decreasing their welfare in the long-run.[88]

Morevoer, there is no direct, obvious mapping between the outputs of a system and the benefits it creates.[89] Instead, the benefits depend on the users, on their perceptions of the outputs in their own context,[90] and on how the outputs impact them.[91] Parity might be more harmful for certain populations than others.

By equalizing an error rate between groups, the disadvantaged groups for which detrimental errors are made might have less time and ability to ask for recourse over erroneous decisions.[92]

> **Example: Job recommendation - Gap between system's outputs and outcomes of the inference subjects.**
> Hiring and job recommendation platforms send pre-employment application tests to a selected batch of candidates. They consider their functioning fair by giving equal opportunities (i.e. an equal percentage of tests) to job seekers of different demographics or of legally-protected groups such as people with disabilities. However, they might actually disadvantage the protected groups: blind job seekers would not be able to correctly fill in these online tests.

Besides, equalising an output distribution across groups does not mean that the outcomes within the groups are fair.

> **Example: College admissions - fairness metrics and causes.** While a similar percentage of men and women may be admitted to a university, it might be that the admitted men all have higher qualifications than the rejected men, while the admitted women might have inconsistent qualifications with some highly-qualified women unfairly rejected and some less qualified women admitted.
>
> In such case, having output parity would not necessarily be considered fair for the women candidates who would expect to be admitted if they have higher qualifications than other admitted women.

This is why individual fairness metrics which focus on the similarities between individuals while ignoring the protected attributes have been proposed.

However, these metrics are also limited, firstly due to the subjectivity and difficulty in defining what similarity means for various use-cases.

---

87 **Alan Lundgard. 2020.** Measuring Justice in Machine Learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 680. https://doi.org/10.1145/3351095.3372838

88 **Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2019.** Delayed impact of fair machine learning. In Proceedings of the 28th International Joint Conference on Artificial Intelligence. AAAI Press, 6196– 6200. https://arxiv.org/abs/1803.04383

89 **Milli, Smitha, et al.** "The social cost of strategic classification." Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019.

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

91 **Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021.** Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application 8 (2021).

92 **Anna Lauren Hoffmann. 2019.** Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society 22, 7 (2019), 900–915.; Milli, Smitha, et al. "The social cost of strategic classification." Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019.

Besides, similar individuals, even though treated similarly, might all be treated in unjustified ways.[93]

For instance, all highly qualified university candidates having studied a specific field could be rejected, while the individual fairness metric would return a fair measurement.

Furthermore, arguing for having fair models with individual fairness metrics where the similarity measure does not include protected attributes, implicitly assumes that it is equally easy for different groups to obtain the same output.[94] Yet, this assumption is often wrong due to the existence of structural disadvantages.

2.1.2 Mutually exclusive notions of fairness

Within a single application, different stakeholders might deem as important different notions of parity.

However, parity notions are shown to be mutually exclusive in the machine learning setup (impossibility theorems say that multiple fairness metrics cannot get high measures simultaneously in a machine learning model), due to the statistical functioning of machine learning models and the unavoidable inference errors it leads to.[95] This forces the need to choose to prioritize one stakeholder.

> **Example: Recidivism prediction - Impossibility theorem.** In the recidivism example, we discussed multiple metrics and what they mean for different stakeholders, especially the decision-maker, an innocent defendant and society (chapter A subsubsection 3.1.1).
>
> The impossibility theorem in this case stipulates that only one of the parity metrics adhering to the interests of one of the stakeholders can be reached, while the others statistically cannot. Concretely, this means that the inferences of the machine learning model will never be considered fair for all stakeholders.

When multiple metrics are considered important, due to the impossibility theorems, either the requirements of the system should be revised, or one needs to accept that it is not possible to fulfil the requirements for fair outcomes in an automated manner and the deployment of the system needs to be questioned.

Reuben Binns also argues that the impossibility theorems are due to different conceptualisations of the world: either that observed output differences are due to unfair inequalities, or to individual choices.[96]

In the first case, both individual or group metrics that rely on error rates or task-relevant features can be selected as the data to reflect a desirable case. In the second case, any metric to check equality in outputs could appear relevant.

In the current literature, the impossibility theorem is addressed in a simplistic manner. Authors necessarily make a choice on the fairness metric to debias a model.

This means that, from their vantage point, they get to determine the relevant conceptualisation of the world as well as the trade-offs with the other notions of fairness and accuracy relevant to the model. This choice unavoidably biases the model towards harmful outcomes for certain populations, to the benefit of others.

This decision about the requirements and/or the non-implementation of the system should not be up to the technologists alone, especially given its societal implications. Instead, individuals or institutions who are more aware of the context in which the system will be deployed could possibly make an informed judgement.

Having the developers or owners of the systems make a decision also means that a centralised authority chooses which metric is potentially closest to a just outcome. Thus, there is an

accumulation of power, a reduction of discretion for people in legal institutions, and most likely no accountable process to decide and to contest which metric might or might not be appropriate.

2.1.3 The questionable definition of protected attributes

Most fairness metrics and subsequent debiasing methods rely on a definition of protected attributes. However, the act of defining protected attributes and the values they can take is reductive and harmful.

Certain attributes cannot be reduced to a simple fixed vector as their conception might be more complex, possibly ambiguous and with multiple definitions. For instance, race attributes in existing data reflect only a few aspects of the multidimensional concept of race.[97]

The ways in which the values of an attribute are defined (e.g. gender as a binary concept) might ignore certain populations completely, or force individuals into non-representative values. Besides, the phenomenon an attribute is expected to reflect might not necessarily be fixed in time, location or context, e.g. notions of gender or age might change over time depending on how a person identifies at different moments, and might be multidimensional in nature.[98]

However, current data schema and data management infrastructures for the datasets do not support the multidimensionality and the flexibility of the concepts (e.g. once the data is collected, it is not easily modifiable anymore).

In turn, when ill-defined attributes are used for bias assessments or debiasing, an incorrect or incomplete notion of bias is tackled. For instance, a system might seem not to be gender biased according to one definition of the protected attribute gender, but this definition might be missing certain values (e.g. non-binary genders), which, if included, could lead to a different conclusion.

Debiasing does not address any of the considerations around protected attributes mentioned above, but relies on these to make any computation.

---

93 **Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. 2019.** This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 1–36.

94 **Binns, Reuben.** "On the apparent conflict between individual and group fairness." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.

95 **Alexandra Chouldechova. 2017**. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data 5, 2 (2017), 153–163.; Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

96 **Binns, Reuben.** "On the apparent conflict between individual and group fairness." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020. https://arxiv.org/abs/1912.06883

97 **Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020.** Towards a critical race methodology in algorithmic fairness. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 501–512. https://arxiv.org/abs/1912.03593

98 **Bonnie Ruberg and Spencer Ruelos. 2020.** Data for queer lives: How LGBTQ gender and sexuality identities challenge norms of demographics. Big Data & Society 7, 1 (2020), 2053951720933286. https://journals.sagepub.com/doi/full/10.1177/2053951720933286

2.2 A system's view of discrimination

When enlarging our view of machine learning models from their outputs to how these outputs are used in practice by different stakeholders, we identify a further misalignment between actual discrimination issues and their conceptualisations in computer science. This misalignment is a real obstacle to ensuring non-discrimination in practice.

In many cases where parity could seem fair in theory, its realisation fails to account for the whole context. The machine learning setup on which parity is verified is insensitive to the decisions individuals actually make based on the outputs, and to the specificities of individuals for which these unbiased systems are actually not beneficial.[99]

2.2.1 The misalignment between system's outcome and decisions

Let us assume the use of an automated decision-making system with a human in the loop. Contrary to the assumption that fairness metrics make, the user of the system does not necessarily take the decision suggested by the system's output.[100]

For instance, not all judges follow the recommendations of recidivism prediction models, and not all doctors follow the diagnostics outputted by XRay-based disease classifiers, since they do not all trust the systems in similar ways.

Consequently, the predictions outputted by the model might be considered unbiased according to certain metrics, but the following human decisions could be biased.[101] Conversely, claiming that a system is unfair due to biased outputs is not always adapted since the final human decisions might re-establish "fairness".

One might consequently want to continuously monitor human decisions to ensure unbiased decisions according to specified metrics. Yet, this would be a difficult process due to the constraints it would impose and due to its surveillance implications.

2.2.2 The limited impact of debiasing on causes of discrimination

Equally, re-allocating a resource often fails to address the causes of the inequalities. It might serve as a satisfying patch for discrimination in the short-term, but it might also reinforce certain harms that cannot be formalised with fairness metrics.[102]

Example: College admissions - biased outcomes or causes. Fazelpour and Lipton take the example of college admissions in the US, where students of different sensitive groups are disproportionately represented for various reasons (including historical and institutional discrimination).[103]

Debiasing methods would enforce equal admissions for all groups. However, they might reinforce existing biases such as gender stereotypes. They might identify women based on certain subfields that they are more likely to choose and, therefore, increase the number of women in these subfields specifically to achieve admission parity, while keeping a lower number of women in the subfields where they are already a minority – whereas more might apply recently.

We believe that more holistic approaches should be welcomed to address the underlying structural causes of these issues. Technical ones can only propose re-allocations of systems' outputs, which are only the surfaces of the issues.

2.2.3 Discrimination short of intersectionality

In order to analyse biases in the case of intersectional discrimination,[104] researchers and practitioners employ the same group of fairness metrics discussed in the rest of the report. For that, the different protected attributes that form the intersectional issues are simply combined into a single attribute with which a protected and a non-protected group can be defined, e.g. gender and race would be the two axes of discrimination, which would be collapsed into a single attribute whose values indicate different permutations of gender and race in the dataset.

This approach fails to address the complexity of the intersectional forms of discrimination people face in the environment of the system.[105] Moreover, there are cases where multiple groups defined over various protected attributes should be considered simultaneously.

However, group fairness metrics only allow analysts to compare two groups at a time, which becomes a limitation for doing an intersectional analysis.[106]

The works by Wachter and Kearns are useful examples for pointing to the limitations of debiasing methods.[107]

Both papers attempt to account for intersectionality by considering differences in outputs for subgroups made up of combinations of attributes. In this case, all permutations of any two groups can be tested to identify potential discrimination.

Wachter considers subgroups of Asian, Black and White women, and says that the fairness metric it proposes, "provides the necessary statistical evidence to compare the magnitude of outcomes and potential disparity between all affected protected groups (for which data is available in a given case). It is thus a tool that enables identification and assessment of potential discrimination, but does not aim to make normative, fundamentally political case-specific determinations normally reserved for judicial interpretation, such as who is a legitimate comparator group or what is an appropriate threshold for illegal disparity in a given case."

This means that a subgroup of interest would be compared to any other possible subgroup to identify an issue of intersectional discrimination, or to subgroups that a judge would deem relevant. Kearns proposes identifying "exponentially many subgroups" ("a combinatorially large or even infinite collection of structured subgroups definable over protected attributes.").

---

99 **Sam Corbett-Davies and Sharad Goel. 2018.** The measure and mismeasure of fairness: A critical review of fair machine learning. (2018).; Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020. POTs: protective optimization technologies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 177–188.; Alan Lundgard. 2020. Measuring Justice in Machine Learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 680. https://doi.org/10.1145/3351095.3372838; Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application 8 (2021).

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

101 **Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019.** Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency. 59–68. https://dl.acm.org/doi/10.1145/3287560.3287598

But this would be problematic. It is computationally hard to perform audits using such metrics, and it could lead to the post-hoc creation of groups based on individuals who do or do not receive positive labels, in order to trick the metrics into showing fair measures.

Intersectionality, originally developed to expose the specific ways in which the discrimination of Black women in in stitutions and social relations are not recognised, is not so much about belonging to a subgroup which receives different outputs or less correct outputs than other subgroups.

A large body of theory and empirical studies identifies the different and complex modes of discrimination that threaten people who sit at the intersection of different oppressed groups. They expose the ways in which intersectional discrimination is produced in interactions and is socially contingent.[108]

By treating intersectionality as a comparison of subgroup outputs (e.g. Wachter provides, "measurements for making comparisons across protected groups in terms of the distribution of outcomes"), the complex manifestations of intersectional discrimination are flattened out, and the possibility to contest them is eliminated.

This goes against the gist of intersectionality and the way the body of work and activism aims to expose the limitations of our typical understandings of discrimination. It surfaces societal assumptions that have also informed our institutional practices, for example, on how thinking about race is mainly imagined through the experiences of Black men, and how considerations of gender are mainly understood through the experiences of White women.

The many works on the subject try to show that these dominant understandings of discrimination themselves make it hard to grasp the very different ways Black women, lesbians with a Maghrebi background, etc. experience discrimination in social situations or institutions.

By misunderstanding intersectionality solely as the membership of subgroups, bias metrics ironically stipulate exactly that what intersectionality intends to dispute: that discrimination is one and the same for all.

2.2.4 The erasure of broader externalities

Until now, we have focused on the direct issues in the outputs of the machine learning models and their users.

Yet, a system is made up of the "machine" in which the models are integrated and an environment in which this machine is deployed.[109] Whilst debiasing does not account for the broader environment (except the end-users of the "machine"), this environment can nonetheless also be negatively impacted.

Selbst et al. highlight that introducing a technology into an environment necessarily impacts the initial environment, its organisation and possibly its values.[110] Verifying that a system is fair with the current focus on models' outputs is, then, not enough, as we also need to analyse the negative impact the new system might have on the entire, original environment – this is what they term the ripple effect.

In particular, the fairness metrics create "unbiased" systems for the "end-users" of the models (i.e. the inputs of the models). Doing so, they leave out other stakeholders and entities in the environment of the systems, that can also be indirectly impacted by the models and the debiasing methods. Particularly, various negative externalities remain unconsidered in bias and fairness frameworks.

Example: Negative externalities left out of the bias frame. Routing applications might fairly route their users (e.g. each of them have similar travel time), while neglecting other issues caused by the

applications such as congestion and damages on roads that are often recommended.[111] Self-regulated housing markets, while not actively discriminating against any of their users, negatively impact the neighbourhoods in which the housing platforms are not implanted.[112]

2.2.5 Aspects of fairness left out from debiasing

Debiasing also leaves out of consideration reasons for discrimination that cannot be scrutinised through the outputs of a machine learning system. Discrimination is often seen in law and policy in terms of "motive, evidenced intent of exclusion, and causality, rather than simply outcomes".[113]

However, none of the fairness metrics used for debiasing accounts for these notions. Instead, they all rely on comparing outputs, and we cannot say that machine learning models have intents per se.[114] Besides, procedural justice encourages analysing existing decision structures. Instead, debiasing solely focuses on the outputs of the models and does not reflect the context of application where the models are employed, the possibility of recourse and explanations around decisions, etc.

For instance, one might achieve fairness in a facial recognition system, which could be desired for simple applications such as for unlocking laptops and phones. Yet, this system might be used by the police to target subpopulations more efficiently, in which case having the technology adhering to any fairness notion itself causes the problem.[115]

Similarly, uses of AI at the border might be deemed biased according to the application of certain fairness metrics (e.g. the system is not as accurate at identifying people from one ethnicity as the other), and could possibly be debiased with regard to these metrics. Yet, they might remain harmful regarding a very different type of issue as they are discriminatory by nature and in practice.[116]

Reuben Binns also argues that existing fairness definitions do not account for the principle of, 'individual justice'; the idea that individuals should be assessed on their own qualities, circumstances, and attributes, not on the basis of generalisations about groups of which they happen to be a member."[117]

Indeed, machine learning models make inferences on people based on a set of features that only partially describes them (and allows to compare them one to the other).

They do not allow for making decisions based solely on the individual, "disregarding any previous knowledge that may have been inferred from previous similar cases, and potentially incorporating new kinds of information and reasoning particular to the case."

---

102 **Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. 2019.** This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 1–36. https://arxiv.org/abs/1909.11869

103 **Sina Fazelpour and Zachary C. Lipton. 2020.** Algorithmic Fairness from a Non-Ideal Perspective. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (New York, NY, USA) (AIES '20). Association for Computing Machinery, New York, NY, USA, 57 – 63. https://arxiv.org/abs/2001.09773

104 **Javier Sanchez-Monedero, Lina Dencik, and Lilian Edwards. 2020.** What does it mean to 'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 458–468. https://arxiv.org/abs/1910.06144

105 **Anna Lauren Hoffmann. 2019.** Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society 22, 7 (2019), 900–915. https://www.tandfonline.com/doi/abs/10.1080/1369118X.2019.1573912?journalCode=rics20

106 **Javier Sanchez-Monedero, Lina Dencik, and Lilian Edwards. 2020.** What does it mean to 'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 458–468.

107 **Wachter, Sandra, Brent Mittelstadt, and Chris Russell.** "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI." Computer Law & Security Review 41 (2021): 105567.; Kearns, Michael, et al. "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness." International Conference on Machine Learning. Pmachine learningR, 2018. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3547922

2.3 Policy implications

All the identified shortcomings expose a lack of consideration for the diverse contexts in which the ADMs are employed, and instead favour easy-to-quantify, one-size-fits-all algorithmic measures that claim to serve the mitigation of what a few people decide to deem unfair.

These shortcomings hint at the impossibility of remedying discriminatory effects of AI with a purely technical approach, as conceptualisations typically remain limited. This limitation is due to both the bluntness of fairness metrics, as well as the exclusion of the ultimate goal and environmental impact of the system from the debiasing analysis.

The separation of the system goal and environmental impact from the debiasing approaches means that a system that distributes bad outcomes evenly can be considered "unbiased".[118]

That debiasing approaches empower the service providers to decide what counts as harms and which measure of debiasing is sufficient raises further concerns about promoting debiasing, especially as a way to attend to the needs of already marginalised communities.

Whilst all policy documents start from the premise that algorithms may be harmful, they fail to reflect on whether debiasing approaches are free of harms or how their shortcomings may further damage vulnerable populations.

None of the policy documents discuss the simplifications that the techo-centric view of debiasing imposes, which means they also do not propose strategies that complement technical approaches.

In light of the limitations we expose above, we hope that policymakers and other relevant actors can better situate debiasing approaches. Going forward (see chapter E), the proposals to use debiasing approaches for "correcting" discriminatory effects of AI should be moderated in light of these limitations.

Advocates should insist that the evaluation of inequalities brought about by AI include technical and non-technical accounts that take into consideration the complexity of the context of its application.

108 **Anna Lauren Hoffmann. 2019.** Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society 22, 7 (2019), 900–915.

109 **Michael Jackson. 1995.** The world and the machine. In 1995 17th International Conference on Software Engineering. IEEE, 283–283. http://mcs.open.ac.uk/mj665/icse17kn.pdf

110 **Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019.** Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency. 59–68. https://dl.acm.org/doi/10.1145/3287560.3287598

111 **Rebekah Overdorf, Bogdan Kulynych, Ero Balsa, Carmela Troncoso, and Seda Gürses. 2018.** Questioning the assumptions behind fairness solutions.

112 **Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020.** POTs: protective optimization technologies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 177–188. https://arxiv.org/abs/1806.02711

113 **Alice Xiang and Inioluwa Deborah Raji. 2019.** On the Legal Compatibility of Fairness Definitions.

114 **Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019.** Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency. 59–68.

3. On the limitations of debiasing methods in practice

In the first part of this chapter, we identified the theoretical limitations of the bias framing for addressing discrimination due to the outputs of machine learning systems.

One might be tempted to oppose these theoretical considerations, by arguing that it would also be impossible for a non-data-driven and non-automated decision process, i.e. a decision making process conducted by humans, to be entirely unbiased.

We will discuss problems with this comparison from a political economic lens in Part D.
For now, let us assume this is a reasonable characterisation and there is a context in which debiasing approaches would merit application.
What challenges are posed then?

Since debiasing can be seen as a purely technical proposition to discrimination, and technical tools, especially machine learning ones, can never entirely achieve their objective, it is reasonable to ask how well, how efficiently and how effectively debiasing addresses the issues for which it is made.

How usable are these tools, and how feasible are their applications in practice? In case of a gap with expectations, to what extent does this gap affect the initial objectives? Given potential discriminatory effects and limitations of debiasing, how can we assess whether auto-mated decision systems still are desirable in a given domain?

We tackle these questions and show that many obstacles render the application of debiasing approaches questionable in practice. These questions arise both in the ability of these methods to mitigate biases, and in terms of being able to assess the successful application of these methods (e.g. through audits).

In particular, we explain that debiasing methods are limited in performance due to the statistical properties of machine learning, and that applying debiasing methods or even auditing for bias in practice raises practical challenges (e.g. definition and collection of representative datasets) that make the task potentially unfeasible.

115 **European Digital Rights. 2020.** Ban Biometric Mass Surveillance https://edri.org/wp-content/uploads/2020/05/Paper-Ban-Biometric-Mass-Surveillance.pdf

116 **Petra Molnar 2020.** Technological Testing Grounds: Migration Management and Reflections from the Ground Up. (published by EDRi) https://edri.org/our-work/technological-testing-grounds-border-tech-is-experimenting-with-peoples-lives

117 **Binns, Reuben.** "On the apparent conflict between individual and group fairness." Proceedings of the 2020 conference on fairness, accountability, and transparency.

118 **Keyes, Os, Jevan Hutson, and Meredith Durbin.** "A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry." Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. https://ironholds.org/resources/papers/mulching.pdf

3.1 The perfomance limitations of debiasing methods

Even under ideal conditions where any practical issue would be resolved, debiasing methods exhibit limitations in their performance. They do not necessarily allow for a fully "unbiased" model, and they often come at the expense of a model's accuracy.

3.1.1 The statistical nature of machine learning

As we mentioned earlier, there are necessary trade-offs to be made between various metrics, not only with regard to fairness metrics but also to performance metrics.

For instance, fairness through unawareness consists of removing protected attributes and their proxy attributes from a dataset. It has been shown to not achieve high accuracy performance, and high fairness for most fairness metrics,[119] due to the limited information available within a dataset, and the limited control given by not having the protected attributes available.

It is statistically impossible in many scenarios to have both an entirely fair and accurate model.

While this can be due to incorrect datasets that do not perfectly reflect the expected outputs or the diversity of population, it is also often due to the statistical nature of machine learning algorithms.

The way machine learning algorithms function imposes a trade-off between the diversity of data patterns to learn, and the complexity of the selected algorithm. It is not necessarily because a dataset is larger that an algorithm can be trained into a more accurate model, as there is only a limited amount of information an algorithm can learn (this is the bias-variance trade-off that the machine learning community faces).

3.1.2 The dependencies within the machine learning pipeline

Biases do not solely arise from the datasets or single activities of the machine learning pipeline (e.g. data processing, model training), but from combinations of activities.

Hence, debiasing a model through one method is often not sufficient to obtain so-called unbiased predictions. Biases might reside in small proportions in a dataset, but be largely amplified by the subsequent choices made when designing and developing a machine learning algorithm.

The choice of features might remove necessary information to apply a debiasing method within an algorithm or when post-processing. The choice of optimisation objective for the algorithm and the choice of error measure might contradict the fairness objectives.

Post-processing a machine learning model to respect additional constraints such as with pruning and quantisation techniques of deep neural networks for memory, latency or energy constraints has also been shown to impact greatly the fairness in the outputs of a model.[120]

Besides, as we show in the next chapter, the initial way in which the problem is defined might already raise harms before even talking about data or models.

Systems might also appear unbiased in development but be revealed as biased when deployed on the new data inputted to the system in deployment.

Yet, there exists no principled method to deal with such biases arising. Such biases are due to differences in data distributions between development and deployment time (data shifts), that can arise for multiple reasons. The populations on which the models are applied might simply change over time.

The data engineering pipelines themselves might also differ between training and deployment due to external constraints, making the data inputted to a model different from the training ones.

For example, a government might install a data capture set-up (e.g. cameras or other sensors) to perform facial recognition, which is different from the one used to capture the training dataset, for practical, cost, or scale reasons.

Besides, what the model is expected to infer might also change over time, due to changes in the ways humans think and behave (concept drift). These changes would potentially decrease the accuracy and fairness of the system's inferences.[121]

These considerations suggest that addressing biases due to various activities, and to seemingly insignificant choices of parameters for these activities in development and deployment, may require infrastructure and process changes.

Practitioners and researchers are currently far from having common machine learning processes. This means that different processes used in different systems may introduce biases in unexpected ways, a matter that is currently hard to evaluate. Whether standardisation of the pipelines or evaluation methods may improve the detection or the mitigation of such biases is an open research question.

Individual AI components are also regularly adapted to be used in systems different from the ones they were initially created for. Such adaptation can raise new bias issues that are not necessarily accounted for in scientific literature and policies.

Machine learning models, and more particularly deep learning ones, require large amounts of data to be trained and tested on, and massive computational power to train models. Hence, it is often more cost-efficient to re-use existing datasets to pre-train a model (or to use an already pre-trained model), before fine-tuning it with a new dataset for a specific purpose using transfer learning.

Even though the original dataset might not be considered biased for its initial application, fine-tuning a model on another dataset might raise new biases, due to the new dataset and its interaction with the transfer learning process.

Hence, it is not sufficient to produce technically "unbiased" datasets or "unbiased" pre-trained models; attention should also be given to the new models that are based on these components.

---

119 **Kusner, Matt, et al.** "Counterfactual fairness." Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. https://arxiv.org/abs/1703.06856

120 **Hooker, Sara, et al.** "Characterising bias in compressed models." https://arxiv.org/abs/2010.03058

121 **Singh, Harvineet, et al.** "Fairness violations and mitigation under covariate shift." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021. https://arxiv.org/abs/1911.00677

3.2 Practical challenges in setting up auditing and debiasing methods

Debiasing and bias auditing both require choosing a fairness metric, and then either applying the metric to the outputs of a model for evaluation, or applying a corresponding debiasing method for making the outputs "fairer".

In practice, it is often challenging to proceed in each of these activities. This can be due to the difficulty in translating contextual information about discrimination into a formal metric.

It can also be due to the difficulties in accessing relevant data about a specific machine learning system, or about specific individuals to apply the metrics or debiasing methods. We describe these challenges further in the next paragraphs.

3.2.1 Challenges in setting up relevant metrics

Anticipating harmed populations. A first socio-technical challenge that applies both to auditing and debiasing a system is to identify the populations or individuals which are (and/or should not be) impacted by biases in a system.

Who could be imaginative enough to predict all kinds of harms that a system might introduce when deployed in an environment? Is it even possible to do so? Industry practitioners stress the difficulty in envisioning all harms to audit, and all populations to consider before deploying a system.[122]

This is challenging as the harms of machine learning are application and context dependent, and cannot be defined universally. Although laws mention certain protected attributes, these are not sufficient to envision all possible cases of discrimination.[123]

Practitioners are usually not domain experts but technical experts, making analysis of the context an awkward, if not impossible, task for them.

In many cases, the process functions backwards. The potential harms are identified after the system is deployed, and after the system has negatively impacted certain populations, and this will be reported to the service provider.

In that sense, debiasing is often an ex-post remedy, a sort of damage control, and, therefore, not an appropriate solution for mitigating all harms.

> **Example: Chatbots - Anticipating harms.** There are many examples of this, such as the Microsoft chatbot whose outputs "became racist" in less than a day on Twitter,[124] the Google vision API which offensively classified Black people as gorillas,[125] Flickr which classified concentration camp images as leisure parks, etc.[126]
>
> Raji et al. mention for instance that transgender Uber drivers have not been able to log onto the application as the facial recognition models did not perform well for them, but this issue was not identified by bias audits.[127]

Translating harms into metrics. Both auditing and debiasing require choosing fairness metrics. However, it is difficult to translate the meaning of these metrics into their concrete impact in the environment of the systems.

That is not only difficult for practitioners who do not know about the environment, but also for domain experts who might not know about the metrics. Besides, the simplifications the metrics make, as explained in the previous section C.2, bring further confusion.

One non-technical solution would be to involve more socially diverse development teams to develop the systems and write their requirements, and to discuss with end-users and relevant stakeholders before deployment.

Such measures could help teams to envision more harms during development and translate these into metrics. However, these solutions remain limited and take place in the presence of power asymmetries between those building systems and those impacted by them.

For example, Holstein et al. interviewed industry practitioners, who indicated that despite running user studies to identify potential harms before deployment, many issues are reported afterwards.[128]

3.2.2 The creation of representative datasets for auditing

Auditing a system for bias requires having a dataset with data samples, on which the system should infer labels, the system's inferences, and possibly the labels reflecting the system's objectives.

But where can one, be it an internal or external auditor, or a developer, find or collect such data samples and inferences? What samples should actually be collected? These questions cannot always be answered in a satisfactory manner, as we explain below, leading to various shortcomings in the auditing phase.

The dataset used for auditing should be representative of the diversity of the populations on which the system is used,[129] and each population should be sufficiently represented in its entire diversity, in order to avoid any misleading fairness assessment.

> **Example: Recidivism prediction - Difficulty in defining a representative set of data samples for auditing.** Let's imagine that we want to audit the recidivism prediction system. If our evaluation dataset contains solely White people and the system usually makes correct inferences for this population but not for the Black population, the disparity would not be raised (assuming a fairness metric based on error rates).
>
> If the dataset contains only a few Black people, and they all have profiles for which a positive outcome is likely (i.e. not re-offending), then the bias assessment would be misled into thinking that both White and Black people get positive outcomes at rather high and similar rates.
>
> The system would be considered fair (assuming a fairness metric based on predicted outcome, e.g. disparate impact). Many more scenarios can exhibit a bias measurement that is strongly misleading.

However, representativeness might be difficult to realise in practice, for multiple reasons.

Accessing relevant information

Auditors, or even developers, either internal or external to the system's creators, might lack information about the individuals on which the system makes inferences (the inference subjects), especially because the pool of inference subjects might evolve over time.[130] Hence, it is complicated to know the kind of samples to collect or create.

Auditors might also not know about the target labels for the machine learning model, i.e. they might not know about the nature of the predictions the model makes, especially if these change over time (concept drift).

For instance, what the COMPAS model was inferring seemed unclear for ProPublica, which wanted to audit it. Was the model inferring whether someone is likely to recidivate over two years or ten years (based on prior recidivism data), or was it inferring the type of sentence a human judge would have deemed the most adequate (based on prior sentences data)? Auditing would require an informed guess, which might not necessarily be accurate.

Finding relevant data

Assuming that representativeness of a dataset can be ascertained, where can the data samples and labels be found next? This question raises more technical challenges.

Auditing requires using the existing machine learning model to collect its inferences on specific data samples, or to modify its functioning. However, access to the model might not be granted.

Besides, the auditors or developers might need to get familiar with the entire code base of the system to run the model and the adequate data pipelines. While black-box audits (only the outputs can be accessed) is possible, it opens further potential for gaming the audits, as we discuss further in the next section.

Both auditors and developers might be in a situation where they need to collect additional data samples, which can be challenging.
For example, there are naturally less data readily available representing minority populations.

For instance, the datasets, especially the ones used to train machine learning models, might be scraped from the Internet, which is inherently biased as certain populations have easier access to the Internet, and have more data representing them than others. It is hence naturally more difficult to include under-represented minorities in the datasets.[131]

For certain fairness metrics (e.g. individual metrics) however, having a large amount of real data is not necessary for all protected and non-protected groups. Instead, it can be sufficient to generate synthetic data by only varying the protected attribute(s) and observing whether the model's predictions remain the same. There, once again, the main challenge is to identify the protected attributes to monitor.

Besides, the data requesters might not be aware of the exact characteristics of the data they are looking for, and hence might also not know where to look for them. For instance, Holstein et al. interview practitioners who identified that their image captioning system was underperforming on celebrities from certain populations were not able to improve it as they did not know what the "foreign" celebrities looked like.[132]

---

122 **Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019.** Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–16. https://dl.acm.org/doi/10.1145/3290605.3300830

123 **Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019.** Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency. 59–68. https://dl.acm.org/doi/10.1145/3287560.3287598

124 **Elle Hunt. 2016.** Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

125 **Nicolas Kayser-Bril. 2020.** Google apologizes after its Vision AI produced racist results. https://algorithmwatch.org/en/google-vision-racism

126 **Richard Gray. 2015.** Flickr's autotag system mislabels concentration camps as 'jungle gyms'. https://www.dailymail.co.uk/sciencetech/article-3093074/Flickr-s-autotag-turns-offensive-Image-recognition-software-mislabels-concentration-camps-jungle-gyms.html

127 **Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020.** Saving face: Investigating the ethical concerns of facial recognition auditing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 145–151. https://arxiv.org/abs/2001.00964

128 **Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019.** Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–16. https://dl.acm.org/doi/10.1145/3290605.3300830

129 **Chasalow, Kyla, and Karen Levy.** "Representativeness in Statistics, Politics, and Machine Learning." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021. https://arxiv.org/abs/2101.03827

130 **Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020.** Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 469–481. https://arxiv.org/abs/1906.02659

> **Example: Recidivism prediction - Technical difficulties in building datasets.** For recidivism prediction, ProPublica, which audited the COMPAS system, had to build a new dataset by merging various sources of information, and had to get access to the predictions of the actual system.[133] They also had to define the labels the system is expected to infer and annotate the dataset to compare the system's prediction with these labels.
>
> This is an ambiguous task when the service provider does not reveal its training data or its data collection strategy.

Creating synthetic data or labels is also difficult. Due to historical biases, it is challenging to envision so-called "unbiased" data labels for the different data samples, especially in cases of developers who might not have domain knowledge.[134]

Accessing sensitive data might be harmful

Various legal constraints might hinder the collection of needed datasets.

Particularly, privacy issues [Kulynych et al. 2020; Raji et al. 2020] might arise, as we further discuss in Chapter D subsection 2.1, especially around protected attributes.[135]

Even if collectable, the ways in which they are measured and transformed into data inputs might not accurately reflect the real notion behind the protected attributes, but instead provide incorrect proxies, leading to models that are seemingly "unbiased", but not in reality.

> **Example: Computer vision - Debiaing protected attributes or proxies.**
> In image-based systems where race information is usually not collected, a proxy used for race is skin colour.[136]
>
> However, there is no one-to-one mapping between skin colour, which is a phenotypical trait, and race, which is a social construct. Hence, applying a debiasing method over skin colour might not actually debias a model over race (assuming that it is important to debias a model along this line).

Besides, certain populations might also generally not provide certain data for several reasons, e.g. overweight people might not communicate their actual weight to insurance companies in cases where they could be discriminated for it.

In this case, collecting such data would be harmful to these people, as the machine learning models trained on this data could make inferences that disadvantage them. Paradoxically, auditing, while aiming at monitoring the fairness of a model's outcomes for unprivileged, often minority populations, raises further harms for them, since collecting more data leads to over-policing minorities and mass-surveillance.

In Europe, the availability of data around protected attributes and minorities is scarce and contested. Data collection on forms of discrimination may vary massively and is generally subject to political contestation, especially when it comes to matters of race and ethnicity,

> "because of the great variety of stakeholders whose consensus it presupposes [...]. Some argue that this data collection essentialises ethnic groups or contributes to race discrimination.
>
> Others are concerned that migration, language, education level and poverty data are not effective proxies for measuring discrimination based on racial and ethnic origin. [...] NGOs often disagree on whether or not collecting data on racial and ethnic origin is desirable.
>
> In general, long standing, dominant, but not necessarily minority-led anti-racist NGOs, Jewish and Roma communities oppose data collection, while groups advocating against hate speech/crimes or

representing non-recognised communities are more vocal supporters."[137]

Given the political nature of the question as to whether and when identity attributes should be collected, and also concerns about how "the way we count" may end up reinforcing categories that are themselves discriminatory and violent (see Chapter C.2.1.3), great care and attention should be practiced in normalising the collection of this data for debiasing or bias auditing.

Dean Spade, a thinker on trans issues and the law, has coined the term "administrative violence" to refer to the way that administrative systems such as the law - run by the state - "create narrow categories of gender and force people into them in order to get their basic needs met", a common example of this kind of violence and normalisation.[138]

Further, if sensitive categories are introduced into systems, we need mechanisms of oversight and community involvement to ensure these systems do not become systems of administrative violence.

3.2.3 The creation of representative datasets for debiasing

Certain debiasing methods also require the collection of additional data samples, often for the datasets to become more representative of the diversity of populations that a model makes inferences on. All of the challenges in collecting data mentioned in the previous subsection about auditing also apply here for debiasing.

131 **Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019.** Does object recognition work for everyone?. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 52–59. https://arxiv.org/abs/1906.02659

132 **Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019.** Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–16. https://dl.acm.org/doi/10.1145/3290605.3300830

133 **Propublica. [n.d.].** How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

134 **Javier S nchez-Monedero, Lina Dencik, and Lilian Edwards. 2020.** What does it mean to 'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 458–468. https://arxiv.org/abs/1910.06144

135 **Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020.** POTs: protective optimization technologies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 177–188.; **Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020.** Saving face: Investigating the ethical concerns of facial recognition auditing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 145–151. https://arxiv.org/abs/1806.02711

136 **Joy Buolamwini and Timnit Gebru. 2018.** Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency. 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

137 **Farkas, Lilla.** "Analysis and comparative review of equality data collection practices in the European Union: Data collection in the field of ethnicity." Directorate-General for Justice and Consumers Directorate D–Equality Unit JUST D 1 (2017).

138 https://ironholds.org/counting-writeup

3.3 Policy implications

Debiasing or bias audits should not become an excuse for service providers to collect sensitive data or to design (administrative) systems using it. As seen in the previous subsections, debiasing and bias auditing are still developing techniques. They suffer from various practical limitations, besides the conceptual ones discussed in Section C.2. We believe that future policy-making should further discuss these limitations, since these are issues that cannot all be solved fully.

Besides, biases can arise in many ways in various places of machine learning systems, e.g. choice of algorithm, differences between the data engineering pipelines at training and deployment time, etc. Yet, most policy documents primarily discuss biases in the data and (implicitly) dataset debiasing, and do not provide guidance to those who apply debiasing approaches.

Particularly, none of the documents mention where bias should be evaluated in the pipeline and when (not in terms of metrics or threshold but pipeline activity).

For instance, it could be investigated in the outputs of the model itself, in the outputs of models that might be working in chain, in the interpretation of the outcomes by potential users, in the way users take actions based on them, etc. The lack of methodological standards or benchmarks in machine learning pipelines is of great concern, especially given the societal implications of using these systems.

We need to encourage research concerning the engineering of pipelines, and how they come to impact machine learning outputs.

Besides, it should be emphasised that one-time debiasing or auditing cannot be sufficient for maintaining fair systems. Instead, data and inferences should be monitored constantly after deployment so as to foresee any changes that could lead to new errors and unfair outcomes. Particularly, this would apply to datasets and pre-trained models that are fine-tuned later.

Finally, policy documents generally argue for applying debiasing and bias auditing without recognising the limitations of applying these methods in practice. Building a relevant dataset and setting up relevant metrics might be impossible, or highly effortful, as industry practitioners argue.[139]

Research is not yet mature enough to provide a precise set of guidelines to support practitioners in doing so, and it might not be able to do so in the future either, due to the case-by-case nature of the work, the difficulties in foreseeing harms and translating them into appropriate metrics, and the challenges in defining and collecting new relevant sets of data while not creating further harms.

More collaboration between decision-makers and policymakers, practitioners, and researchers would be needed to better grasp the challenges of debiasing and bias auditing for each upcoming use-case and application, and for providing more actionable recommendations.

This problem is reinforced by the fact that what is considered a bias in society changes over time, and by the frequent updates done to machine learning models, that would force frequent repetition of the audit process of the models (and their debiasing).[140]

Given the theoretical limitations described in Section C.2 and the practical limitations we have covered here, policymakers should cease to promote "debiasing" as the ultimate solution to the discriminatory impact of AI.

Instead, we recommend that they focus on independent bias audits as an initial screening step in order to check the machine learning model for the most obvious biases that it could integrate, accompanied by a more comprehensive evaluation of the systems the models are integrated in.

---

139 **Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daum III, Miro Dudik, and Hanna Wallach. 2019.** Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–16. https://dl.acm.org/doi/10.1145/3290605.3300830

140 **Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020.** Saving face: Investigating the ethical concerns of facial recognition auditing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 145–151. https://arxiv.org/abs/2001.00964

4. The dependence on service providers

One last hurdle for performing "accurate" audits or "effective" debiasing as envisioned by the technical measures of bias is the willingness of the service providers. Since debiasing and auditing require access to data and models, only willing service providers can grant such access.

The service providers could also easily perform misleading actions when auditing their system, in order to make the outputs of their systems look unbiased. Externally regulating the audits or verifying that debiasing has been performed is challenging, since it is close to impossible to define and collect appropriate datasets for arbitrary use-cases (as explained in section C.3 above).

The recent ban that Facebook imposed on researchers who collected data about the platform in order to study its advertising system illustrates this difficulty.[141]

4.1 The necessary incentives for objective actions

4.1.1 Gaming the audit

Auditors can intentionally make use of the difficulties or simply of the freedom they have to create the audit data-set (since often unknown information is needed for defining the representative dataset) for gaming the results of the audits, or what is disclosed.

There is currently no way to guarantee that the audit is performed objectively as it is a developing science. Examples outside the debiasing context but concerning auditing companies for questions of privacy, and the questioned validity of the reported results, include Google,[142] and Amazon,[143] and illustrate once more the subjectivity of audits, even when conducted by third-parties.

They can reverse-engineer the selected metrics to design a set of dataset characteristics that naturally fit with the metric requirements.

Varying the sampling of the available dataset and the granularity level of its attributes can "trick" the measure. For instance, Sánchez-Monedero et al. studied multiple companies that perform resume screening for hiring and that claim to be unbiased.[144] They showed that some of the companies use simple definitions of protected

attributes (e.g. binary gender) and exclude certain of them (e.g. no analysis of social class or disability discrimination), resulting in the company appearing fair.

Additionally, choosing the fairness metrics and acceptance thresholds, and defining the values of the protected attributes after performing some experiments, can help trick the audit into characterising the system as fair.[145]

The service provider can also share artificial data samples or incorrect labels, simply to match the metrics chosen by the auditors. There might not be any guarantee for the trustworthiness of the data that would be shared.

4.1.2 Hurdles with the service providers

The assumption that service providers have the necessary and sufficient incentives to identify biases and apply debiasing methods does not always hold.[146]

Debiasing might reduce the accuracy of their models, and increase the cost of development (e.g. cost of collecting new training data) as well as the time before deployment. Typically, debiasing literature assumes the just a single service provider is acting at a time. However, harms might be caused by the use of multiple services to build a single system. In such cases, each of the providers could consider setting up debiasing strategies in collaboration, and question of responsibility would arise.

Finally, it might be structurally complex to setup debiasing methods as various steps and components of the machine learning process (e.g. dataset collection and curation pipelines, model development, deployment and update pipelines, etc.) might be handled by different practitioners.

The responsibility for debiasing the systems might then remain unclear. Besides, accountability is difficult to trace back in machine learning settings with various actors, reinforcing the responsibility issue.[147]

These matters are further addressed in Chapter D.2 (production view).

---

141 https://www.theverge.com/2021/8/4/22609020/facebook-bans-academic-researchers-ad-transparency-misinformationnyu-ad-observatory-plug-in

142 https://thehill.com/policy/technology/410568-exclusive-privacy-audit-failed-to-mention-of-google-plus-security-flaw

143 https://www.politico.eu/article/data-at-risk-amazon-security-threat

144 **Javier S nchez-Monedero, Lina Dencik, and Lilian Edwards. 2020.** What does it mean to 'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 458–468. https://arxiv.org/abs/1910.06144

145 **Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020.** Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 469–481. https://arxiv.org/abs/1906.09208

146 **Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020.** POTs: protective optimization technologies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 177–188. https://arxiv.org/abs/1806.02711

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

4.2 Debiasing and auditing at the discretion of the service providers

As auditing or debiasing approaches depend on the service providers, the system values reflected by the bias metrics that auditing and debiasing rely on also depend on these service providers.

Choosing a metric to optimise, debias and audit for is context dependent. There are no "ideal" mappings between machine learning use-cases and fairness metrics. Who should then be responsible for such choice? And how should the threshold of acceptability or rejection of a model based on the metric be set?

For now, service providers make these choices, giving them (and not lawmakers) the discretion to decide what counts as discrimination, when it occurs and what it means to address it sufficiently.

As discussed in the previous subsection, it is challenging for practitioners (or service providers) to choose the metrics.

It was also shown through multiple studies that people with different backgrounds understand fairness metrics differently and have differing opinions on the ones to be prioritized.[148]

Hence, consistent agreement might not be reachable in practice, except if an external public authority is able to make a concrete and informed choice.

But how would such a choice be made, and what are the implications of a process for democratically addressing inequalities?

[148] **Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020.** An empirical study on the perceived fairness of realistic, imperfect machine learning models. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 392–402.; **Debjani Saha, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, and Michael Tschantz. 2020.** Measuring non-expert comprehension of machine learning fairness metrics. In International Conference on Machine Learning. Pmachine learningR, 8377–8387.; **Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2020.** How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. Artificial Intelligence 283 (2020), 103238.; **Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019.** Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2459–2468.; **Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018.** A qualitative exploration of perceptions of algorithmic fairness. In Proceedings of the 2018 CHI conference on human factors in computing systems. https://arxiv.org/abs/1802.01029

4.3 Policy implications

Auditing and debiasing approaches, very much like the rest of AI, entrench centralised control over the enforcement of decisions and policies (of non-discrimination). Policy documents also propose voluntary audits for lower-risk systems.

Yet, structural inequalities and discrimination are actually societal problems that are not solvable from a centralised vantage point.

Doing so concentrates the possibility of negotiation and power in the hands of those who have the centralised control. Debiasing tasks these entities with deciding which inequalities are relevant, and what policies will be applied (in automated systems) to address them.

Given the way in which bigger and better models of machine learning are concentrated in the hands of a few big tech players, this means that the ever more powerful entities that dominate the machine learning market may come to be the arbiters of political problems around discrimination and inequalities.

More generally, machine learning uses a utilitarian logic, the parameters of which are settled from a central vantage point. This empowers machine learning designers and service providers in making decisions that, if successful, come to order the world and shape what counts as utility, as guided by their own political and economic interests.

In the process, applying machine learning transforms social and political questions into technical and economic ones.

How rigorously these technical decisions are made, and attention to their potential societal implications, are hence also determined by the economic, political and organisational limitations under which machine learning is applied.

These challenges necessitate great caution when promoting debiasing approaches as solutions to the inequalities caused by the introduction of AI into an environment.

> **Structural inequalities and discrimination are actually societal problems that are not solvable from a centralised vantage point.**

D.

Alternative framings for AI policymakers

There are many harms due to the integration of AI into digital services that are not captured in the debiasing approach to data and design of algorithms.

The previous sections showed the limitations and contradictions inherent to debiasing approaches in terms of the discriminatory impact they claim to address. Aside from being a burgeoning field that has prematurely been catapulted to front stage by companies and policymakers alike, debiasing approaches carve out only a small portion of the equity harms associated with the introduction of AI.

In particular, by locating potential harms in datasets and machine learning algorithms, debiasing approaches fail to capture the impact of AI more broadly on discrimination and social inequalities.

To avoid trivialising the problem, we believe that policymakers should go beyond a focus on abstract concepts like datasets and algorithms as they pertain to decision making. While these are some of the main abstractions computer scientists use in machine learning research, they do not account fully for the material manifestation of AI in the world.

Similarly, framing systems in terms of automated decision making emphasises a socio-technical view, but leaves out the many ways in which AI is used to produce digital services that may raise similar concerns around social inequalities.

But, if data and algorithms, or ADMs, are not the unit of policy-making, then what is? In order to go beyond seeing AI as a technique in decision making, we sketch alternative views that help to highlight AI's broader impact, with a focus on discriminatory effects and inequalities.

We propose a machine learning and a production view on AI, and sketch additional views, e.g. infrastructural and organisational. In all of these, we bring in a political economic understanding of AI which is absent in most policy documents and is pertinent to matters of equity.

We hope that these framings can help identify new orientations for civil society advocacy and also provide directions for the development of more robust policy-making to address the potential harms of AI.

1. The machine learning view

In the previous chapter, we questioned the model-centric view of debiasing tools. These tools rely on the entities necessary to set up machine learning systems, i.e. algorithms, training data and/or protected attributes and inferences, etc. (presented in Appendix B).

These entities are implicitly presented as unquestionable as they are necessary to the functioning of the technology.

However, can they not themselves also be problematic? Machine learning systems, for instance, use datasets with various attributes describing individuals (e.g. skills, background, etc.) and target decisions (e.g. granting a loan or not), in order to extract data patterns within these.

Implicitly, this assumes that the attributes are relevant to the target decisions, and that new decisions can be made simply by comparing a new individual to individuals in the dataset.

These are strong assumptions which are not discussed debiasing methods, even in though they might also lead to unfairness. Here, we outline such potentially problematic assumptions, that allow us to question the use of machine learning itself in certain contexts.

1.1 Dubious optimisation task definition

It is worth taking a step back from the focus on datasets, models and their biases, and interrogating whether the envisioned task can be performed using machine learning, whether the labels and data that are put forward are scientifically sound for the task, and whether relevant data can actually be found.

AI, especially machine learning, relies on principles that might seem obvious, but that can shape a task in possibly harmful ways. We pinpoint these principles and their issues below.

1.1.1 The principle of reproducing historical data patterns

Machine learning systems performing classification or regression tasks rely on the identification of patterns in training data that reflect past behaviours, in order to learn an inference behaviour.

This inference behaviour is then used to make inferences on new data encountered once the systems are deployed.

The new data samples are "compared" to the training data, and the system infers behaviours that are closest to the ones of the most similar training data, which are used to infer labels for the new samples.

Yet, making inferences by mimicking past behaviours and comparing the new samples that describe the new inference subjects to past training data (generally corresponding to past subjects of a decision) can be harmful in various ways, that are not discussed within the debiasing context.

Implicit repetition of past behaviours

Are the past behaviours desirable, and is it desirable to simply repeat them? This is something to question in the different contexts of application of machine learning. Learning from past behaviours to make new inferences might be harmful in certain cases.

If certain types of population were not encountered in the past, the systems might make irrelevant inferences for them. If the past behaviours were problematic or discriminatory, the new inferences would reproduce problematic behaviours.

> **Example: Hiring recommendations - Repetition of the past.** Raghavan et al. question the idea of using machine learning for job hiring decisions.[149] The machine learning process would inherently skew the task of identifying satisfactory candidates towards finding candidates resembling those who have already been hired, leaving out new, different, qualified candidates that have not been encountered before by the companies.
>
> Fairness metrics would not detect such an issue since the historical training data they are usually applied on would not contain any information on the new types of candidates. Accounting for the new candidates would require a human to foresee all their characteristics, and to possibly build data items representing them and their desired label, before applying the metrics. This would be directly opposed to the machine learning principle of learning patterns in the data and automatically repeating them.

Decision-making by comparing individuals

Do we want to make this decision simply by comparing this individual to others?

This question is implicitly positively answered when choosing to use a machine learning model. Yet, certain notions of justice, for instance, are not comparative, in which case it is not valid to use machine learning to make a decision. However, the prevalence of debiasing frameworks obfuscates such questions in the context of automated decisions.[150]

> **Example: Recidivism prediction - Individual particularities or parity.**
> A system for recidivism prediction could be deemed "unbiased" and legitimate within the bias frameworks if it treats all individuals of two groups similarly.
>
> Yet, before providing similar outputs to different individuals, one question worth asking is the relevance of comparing individuals in order to infer the likelihood of recidivism and the according sentence. Comparing individuals to judge them neglects their individual rights and singularity.
>
> The underlying problems of such comparisons become especially clear when individuals might have been judged by the system based on comparative information that is legally irrelevant or reflective of institutional racism.
>
> The COMPAS Risk Assessment (CORE), for example, relied on questions about separation of parents, friends/acquaintances that have been arrested, as well the impression of the interviewer as to whether the person under arrest is a "gang member".[151] This is also questioned in predictive policing systems in Europe.[152]

1.1.2 Scientific soundness of the system's task or objective

How sound is it to identify patterns between the input data available and the target label? Machine learning relies on the assumption that there exists a relation (formalised as a pattern) between the input data and the target label.

While this relation might not have to be of causal nature, the existence of correlations is also sometimes questionable: is the existence of correlations backed up by prior scientific evidence?

Are we making an assumption that might lead to random and harmful predictions? Such questions are especially timely for image-based tasks.[153]

> **Example: Computer vision - Soundness of the task.** In one of the first large datasets used for image classification, ImageNet, labels such as "orphan" and "professor" have been used, but they do not objectively map to visual properties of someone or something. Is it reasonable to assume that someone's job or orphan status can be inferred from a simple picture?
>
> While it can simply lead to wrong predictions of the models, it can also deviate to harmful practices relating the discredited science of physiognomy, i.e. the idea that someone's character can be identified from their appearance, e.g. gender identity inferred from the face.

A burgeoning critique of machine learning systems and AI has therefore been to fundamentally question the scientific validity of the underlying assumptions and stated objectives of the system, before the examination of the reliability or accuracy of the system.[154]

The increasing reliance on pseudo-scientific assumptions for certain systems, including lie detection, emotion detection and biometric categorisations systems necessitates an initial, broader analysis of whether the stated objectives of certain systems are even scientifically valid.

However, we also caution that neither science or academia is protected from accepting ways of categorising and ordering populations that are very much based on power, majority consensus, or colonial histories. These may normalize oppressive beliefs as scientifically valid, disadvantaging, for example, minorities or racialised others, as has been evident in eugenics and phrenology and their reappearance in AI.[155]

Syed Mustafa Ali has highlighted that the scientific racism that was instrumental to colonialism expressed itself in 'sedimented' ways of knowing and being – based on systems of categorisation, classification, and taxonomisation and the ways that these are manifested in practices, artefacts and technologies".[156]

1.1.3 Desirability of the task

Even when the service providers have established a task as sound and its repetitive nature as acceptable, it still remains important for policymakers to ask whether this task is desirable, i.e. whether the creation of an automated decision system would indeed automate a desirable task, or whether it serves to obscure a questionable one.

---

149 **Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020.** Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 469–481. https://arxiv.org/abs/1906.09208

150 **Sina Fazelpour and Zachary C. Lipton. 2020.** Algorithmic Fairness from a Non-Ideal Perspective. In Proceedings of the AAAI/ ACM Conference on AI, Ethics, and Society (New York, NY, USA) (AIES '20). Association for Computing Machinery, New York , NY , USA , 57 – 63. https://doi.org/10.1145/3375627.3375828

151 **Northpointe. 2011.** Compas Risk Assessment CORE. https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.htm. https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.htm

152 **Williams, Patrick, and Eric Kind.** "Data-driven Policing: The hardwiring of discriminatory policing practices across Europe." (2019). https://www.enar-eu.org/IMG/pdf/data-driven-profiling-web-final.pdf

153 **Kate Crawford and Trevor Paglen. 2019.** Excavating AI: The politics of images in machine learning training sets. Excavating AI (2019).; Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 547–558.

154 **Erikson and Lacerda. 2008.** Charlatanry in forensic speech science, The International Journal of Speech, Language and the Law [IJSSL], vol 14.2, 169–193. https://journals.equinoxpub.com/IJSLL/article/view/3775

Assuming the outputs a system is asked to allocate are undesirable, simply applying debiasing in order to allocate them equitably would not be helpful. Let us imagine a system which would equally allocate bad working conditions to different job seekers. While being fair for all its users, it would also be harmful as it would be allocating negative resources.[157]

Mitchell et al. discuss that different stakeholders might in any case disagree with the overarching goal of the system, hence even with any bias mitigation method, certain stakeholders would not find the system acceptable.[158]

1.1.4 Discretisation of the environment into categories

Machine learning relies on the use of structured data - data organised in tables with multiple attributes and sets of discrete values that each attribute can take.

Hence, machine learning discretises the world in ways which might miss certain categories (if the world can truly be objectively categorized), misrepresent them, and reflect non-universal constructs, such as discretising humans into the two gender categories "male" and "female".

This discretisation activity is highly subjective while objectifying people, is often reflective of embedded power dynamics, and might hence be harmful to populations that do not share the same views on the discretised world.

Also, the sole act of attributing data samples to labels when producing datasets creates one single view of the concept referred to by the label, in turn creating a stereotype for this concept, while negating its diversity of interpretations.[159]

1.1.5 Machine learning's desire for universality

Often, researchers and developers aim at making "universal" machine learning models, i.e. the models' predictions should be accurate on any data that can be found in the world.

For instance, one of the primary goals of computer vision researchers is to allow machines to understand fully any image in the world. This means that the models would be independent of the context in which they are used since any context would be universally included. Economically, developers aim for this universality in order to build service architectures applicable to many contexts with a single model, without having to retrain the model (which is costly, due to the need for data and computational power, and time-consuming) since it should not make errors when data from new contexts appear.

Yet, this ideal desired situation is out of reach due to the technical limitations of machine learning, but also because machine learning imposes representations of the world that are not necessarily satisfying.

Assuming models are universal tends to overlook a number of issues that machine learning is subject to, such as data shifts (as mentioned in Chapter C subsection 3.1.2), and that can create harms especially for vulnerable communities.

---

155 **Birhane, Abeba, and Olivia Guest.** "Towards decolonising computational sciences." https://arxiv.org/abs/2009.14258

156 **Syed Mustafa Ali.** A brief introduction to decolonial computing. XRDS: Crossroads, The ACM Magazine for Students, 22(4):16–21, 2016. https://dl.acm.org/doi/10.1145/2930886

157 **Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020.** POTs: protective optimization technologies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 177–188.

158 **Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021.** Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application 8 (2021).

1.2 Soundness of the data schema design

Once a task is agreed, the machine learning setup imposes the creation of a dataset. What does creating a dataset entail for the desired inference task? To what extent do the ways datasets are formalised reflect the real case? Building a dataset requires defining a set of attributes and discretising the values they can take, and collecting data reflecting such attributes.

These activities impact the inferences made by the machine learning model trained on the data, in ways which can be harmful outside the bias framing.[160]

The definition of attributes and choice of data to populate them is typically not linear but interdependent since attributes are often chosen based on data availability. Hence, the issues we identify in this subsection 1.2 all impact each other.

1.2.1 Problematic definition of attributes

The choice of attributes constituting the dataset impacts how well the model trained on this dataset performs its intended task. An incorrect choice might create various harms.

The choice of incomplete sets of attributes

The selected set of attributes might be incomplete, not providing enough information to properly perform the inference task, such as not providing the amount of a loan one has applied for when predicting one's likelihood to repay it.

Hence, even if the inference task is itself desirable and sound, the attributes chosen themselves might not support the execution of this task. Instead, they might propagate harms either due to random prediction errors or due to spurious and discriminating correlations that the model could have learned from this data.

The choice of irrelevant or problematic attributes

The attributes chosen might simply not be relevant for the task at hand, such as using one's number of siblings to predict whether one is likely to repay a loan. Collecting data on certain attributes might even be considered unfair and possibly illegal.

This can be because they are not the result of volitional decisions, such as using the age or race of an individual to decide on the jail time, contrary to potentially volitional decisions like the number of prior offenses; or because they are privacy-infringing.[161]

The choice of certain attributes for the models might also prevent certain stakeholders from recourse over inferences, such as when the attributes are, "[immutable] (e.g. age 50), conditionally immutable (e.g. has_phd, which can only ≥ change from FALSE -> TRUE), or should not be considered actionable (e.g., married)".[162]

Other attributes do not necessarily have to do with unfairness but with offensiveness.[163] This is especially the case for systems that work with image datasets.

> **Example: Computer vision applications - Offensive labels.** The target labels are often organised in taxonomies, and these taxonomies can reflect offensive constructs.
>
> For instance, ImageNet has in its taxonomy a branch for "bisexual" containing the label "hermaphrodite".

It also supports in the branch "adult body" the binary "male body" and "female body" distinction.

The labels themselves can be pejorative as well [Yang et al. 2020], especially when they are associated to certain data samples, such as the label "kleptomaniac" associated in ImageNet to an image of a woman lying on the beach.

The simplified decision space

The decision space for a decision-maker refers to the choice of target labels.[164]

This choice defines the set of actions or decisions that a decision-maker can take with the help of the corresponding machine learning model.

This choice might greatly impact the environment in which the machine learning model is implemented. For instance, it might reduce the number of possible decisions taken compared to a situation where humans make decisions without any system support.

For instance, loan lending systems often decide either to accept or reject a loan application; and recidivism prediction systems infer whether someone is likely or not to reoffend to decide whether to put or keep them in prison.

A decision-maker could foresee other possibilities, for example, Mitchell et al. argue that a decision maker may consider, "a loan with different interest rates and loan terms"[165]. One could also consider proposing reinsertion programs for the detainees.

One can also imagine some decision makers may want to more fundamentally question the injustices built into the credit system as manifest in subprime loans with very high interest rates, or the criminal justice system as often discussed under the rubric of the carceral industrial complex.

1.2.2 The choice of erroneous data to populate the attributes

While the task could possibly be sound, it might be that the data used in practice are not valid for populating a chosen attribute, for various problematic reasons.

Essentially, either the phenomenon the data should reflect is not measurable or is measurable only with inaccurate proxies, or a satisfactory proxy might exist, but errors might arise from the way this proxy data is collected.

159 **Vinay Uday Prabhu and Abeba Birhane. 2020.** Large image datasets: A pyrrhic win for computer vision?; Dobbe, Roel, Thomas Krendl Gilbert, and Yonatan Mintz. "Hard Choices in Artificial Intelligence."

160 **Solon Barocas and Andrew D Selbst. 2016.** Big data's disparate impact. Calif. L. Rev. 104 (2016), 671.; Abigail Z Jacobs and Hanna Wallach. 2019. Measurement and fairness.

161 **Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018.** Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32.; **Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020.** Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 469–481.

162 **Ustun, Berk, Alexander Spangher, and Yang Liu.** "Actionable recourse in linear classification." Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019.

163 **Kate Crawford and Trevor Paglen. 2019.** Excavating AI: The politics of images in machine learning training sets. Excavating AI (2019). https://www.researchgate.net/publication/352224617_Excavating_AI_the_politics_of_images_in_machine_learning_training_sets

164 **Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021.** Algorithmic Fairness: Choices, As sumptions, and Definitions. Annual Review of Statistics and Its Application 8 (2021).

165 **Mitchell, Shira, et al.** "Algorithmic fairness: Choices, assumptions, and definitions." Annual Review of Statistics and Its Application 8 (2021): 141-163.

The choice of proxy data

The collected data might not truly represent the target attributes - the phenomena being measured - but are proxies for them.

Depending on the nature of the proxy, this can cause various harms. If the proxy is too approximative of the real data or not even scientifically related to the phenomenon, then the machine learning model might learn to perform well solely on these inaccurate data.

This is often the case when it is hard or impossible to collect the needed data as the phenomenon is not measurable easily, if at all.

> **Example: Emotional expression recognition- Validity of available data.**
> The detection of emotional expression is a popular task in machine learning, and has been performed using different types of proxy data for the true, interior, emotional state of an individual. Yet, some of these proxy data, such as facial expression or heart rate, see their relevance contested following existing research on emotions, as their accuracy and suitability for emotion is limited. Using such proxies for performing an inference task would then lead to prediction errors that might be harmful, depending on how the system is used.[166]

The proxy might also be a partial representation of the target data, again often due to difficulties or impossibilities in collecting the real, desired data, as they might not be easily measurable proxies.

> **Example: Recidivism prediction - Validity of available data.**
> Recidivism prediction models are trained on data for recidivism over two years, not over a lifetime, contrary to what the model is aimed at helping a judge decide upon. This is possibly due to the difficulty of finding large amounts of more relevant data. Yet, collecting data over a life-long term would also mean tracking populations indefinitely, surveilling people and reversing the presumption of innocence, which exactly defeat the purpose of using debiasing in order to avoid discrimination.

The incomplete and incorrect collection of data

The data samples that are included in the dataset might reflect an incomplete view of the world due to limitations in the design of the sampling arising from practical reasons or human biases.

For instance, in the recidivism case, only individuals who were released and not in jail and then followed over two years could be included in the datasets, biasing the set of individuals in the training data, as we cannot know accurately what the individuals in jail would have done if they had been released.

Mitchell et al. also mention that the human process leading to the inclusion or not of individuals in the dataset may reflect oppressive social structures (e.g. over-policing of certain minorities).[167]

In relation, the collected data might be wrong due to similar practical constraints. Historical human decisions might indeed be biased (such as for jail time decisions that judges might have made) and consequently can be considered wrong for certain data samples and inference tasks. In such cases, the training data labels collected are flawed from the beginning.

Such sampling and errors raise concerns once the dataset becomes the basis for model training and future decision- making, as it takes away from the discussion and normalises these prior questions.

The erroneous collection of data

Data might not reliably be collected. In cases where data are available but with errors (e.g. missing values for certain attributes, wrong labels, etc.), correcting these errors is a complex task, and might introduce additional biases since the system's developers might not know what the missing data is but infer it from the available majority.

In any case, errors and how they are handled can bias the dataset and raise additional harms in the outputs of the models.

Example: Hiring - Improper collection of data. For a system that performs automatic candidate hiring for a job, the training data would need to contain information about individuals who have already been hired or rejected, how they perform at the job at hand (or another similar one), and their aptitudes for certain tasks, possibly measured using evaluation tests.

However, there might not be enough individuals willing to perform these tests seriously, as these are time-consuming. Taking the tests quickly might lead individuals to enter mistakenly wrong information in the training data, which would train models to make incorrect inferences.

1.3 Policy implications

Machine learning technology comes with its own logic and ontology of the world. For computer scientists, machine learning is a family of techniques that achieve computational outcomes based on empirical data. Machine learning allows its makers a way to categorise and order the world optimally in service of an objective function - a great power with material consequences that debiasing approaches fail to capture.

This machine learning view allows us to capture and understand concerns around the power and legitimacy of developing "representations" of the world using machine learning.[168]

It exposes the ways in which the representations used in machine learning are complicated by the objectives pursued, shortage of scientific rigour or complete lack of scientific validity, dangerous pragmatism in the construction of ontologies, oversimplification, developers' own worldview, and data accessibility constraints.

Table 5 provides an overview of the problems that a machine learning view exposes, and that are not addressed with debiasing approaches.

Going beyond representation, this view supports understanding of how machine learning may enable systems that operate in a way that may

166 **Stark, Luke, and Jesse Hoey.** "The ethics of emotion in artificial intelligence systems." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021. https://dl.acm.org/doi/10.1145/3442188.3445939

167 **Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021.** Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application 8 (2021). https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-042720-125902

disadvantage the ability of certain populations to act or make sense of their social experiences.[169]

Bias discussions do not approach these topics as they do not fit the traditional technical framing of having measurable objectives. Yet, even if a system appears fair/unbiased on the proxy data it is evaluated on, the system might be problematic due to all the issues in the machine learning view.

To highlight such phenomena, Selbst et al. put forward the idea of a solutionism trap: searching for technical solutions to issues coming from technical artifacts, whereas the problem might be socio-technical and cannot be understood solely in terms of technical tools.[170]

It is not only that the technology might not be adapted to find solutions to the goal it is aimed at achieving, but the proposed solutions might at times create or reinforce the issues.

Research on biases has obfuscated these other issues highlighted by the machine learning view to the extent that even activist organisations do not necessarily mention them.

For instance, the moratorium on facial recognition in the US relies on the idea that existing facial recognition software leads to wrongful and biased arrests, yet it questions neither the idea of performing automatic facial recognition itself – except in relation to privacy - nor the way classes/groups are defined in order to measure the biases of the errors.

EDRi has already discussed a subset of the issues from the machine learning view, such as the desirability of the task, the repetition of patterns, concerns related to human rights, etc., but not all of them, (e.g. the (scientific) soundness of the task or of the data used to perform it has not been discussed to the best of our knowledge).

---

168 **Philip Agre. 1997.** Beyond the mirror world: Privacy and the representational practices of computing. Technology and privacy: The new landscape (1997), 29–61.

169 **Syed Mustafa Ali. 2018.** AI and Epistemic Injustice: Whose Knowledge and Authority?. (2018). https://www.academia.edu/37919033/AI_and_Epistemic_Injustice_Chairs_Response_S_M_Ali

170 **Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019.** Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency. 59–68. https://dl.acm.org/doi/10.1145/3287560.3287598

Table 5: Machine learning view: Problems not addressed by debiasing methods in computer science.

Dubious optimization task definition	
Implicit repetition of past behaviours	Machine learning classification and regression tasks rely on the principle of identifiying patterns in data reflecting past behaviours across a set of individuals, which may entrench past practices and override individual harms. E.g. hiring predictions based on previous hires.
(Scicientific) soundness of the inference task	Machine learning assumes a relation between the input data and the target label, e.g., labelling of "orphans" in images.
Desirability of the task	Regardless of fair outcomes, the task itself may be undesirable, e.g., distribution of poor working conditions at scale but fairly.
Discretization of the environment	Machine learning relies on discrete data, while discretization itself is highly subjective, often reflective of embedded power dynamics, and can be harmful when populations do not share the same views, or can be harmed by the discretization of certain categories, e.g., the use of the Fitzpatrick Skin Type for racial disparities reducing race to skin phenotypes.
Machine learning's desire for scale and universality	Machine learning researchers aim at making "universal" models that are accurate on any data, or to develop services that are expected to function across different social contexts. Aside from the impossibility of universality, such an aspiration is likely to result in errors in specific contexts of use, e.g., ImageNet assumes there are universal labels for objects in images, but a same object has a different appearance across the world, leading to certain misclassifications (e.g. kettle, wedding, etc.).
Soundness of the data	
Choice of attributes	Choice of attributes may be incomplete, irrelevant, harmful, and lead to a simplification of the decision space, e.g., ImageNet uses taxonomies that classify sexual minorities as deviant or pathologic.
Choice of data	The data used in a model may be an inappropiate proxy, the dataset may be incomplete or incorrect, and may include erroneous entries impacting the outcome, e.g., the data points used in risk analysis may be based on questions that are based on prejudices, may be incomplete or contain errors. Emotion recognition is often based on facial expression that often does not reflect the inner emotion.

2. The production view

While it is tightly knitted to a scientific field, computing is also a business. AI is typically not just developed for the sake of creating "intelligent machines", for some notion of intelligence, or solely for relieving end users of laborious tasks (e.g. making restaurant reservations with Google Duplex).[171] As a business proposition, AI brings about other considerations that are typically not considered in computer science research.

For example, deploying machine learning systems requires setting up production processes and associated computational infrastructures to collect, process and maintain datasets, as well as to train machine learning models and deploy them.[172] These pipelines and infrastructures, and their production, not only pose hard engineering problems but are deeply shaped by the business logic surrounding them.

However, both in computer science and in most policy-making, the political economic considerations associated with AI are typically abstracted away. They are rarely made explicit when debiasing approaches are discussed, despite the constraints this may pose for the application of debiasing methods.

By considering AI as if it exists independently of the business of computing that underlies its deployment, many of the inequalities arising from the production of AI become invisible.[173]

In the business of computing, machine learning is often promised as a solution to automate, augment and scale costly activities or workflows.[174]

Machine learning is especially shown to be effective when applied to day-to-day operations of an organisation, solving complex resource allocation or logistical problems, or improving production lines in many sectors ranging from manufacturing to creative industries.[175]

This means that many applications of AI will take place in Business to Business (B2B) contexts, and not just in consumer facing (B2C) applications.

In B2B applications, machine learning is considered a viable business proposition as long as it provides either greater or new forms of revenue, or cuts costs. To give an example, we can look at the use of AI chatbots for customer service. Chatbots can be deployed to cut costs by aiding customers in solving their own problems.

A successful chatbot is one that can keep customers from contacting a call-centre, reducing the cost that can accrue with each call.

In the context of chatbots, a policy approach narrowly focused on debiasing would aim to provide services to customers from different subpopulations equally, assuming the only harm of interest is that of fairness in market services.

However, here, the debiasing view leaves out considerable factors driving inequalities between different populations and organisations implicated in production processes. More and more institutions delegate their fundamental operations to scaled-up AI-services and hence AI service providers, for whom profitability depends on the externalisation of costs of contextual needs, failures, or damages, by design, to others.

For example, when chatbot services are adopted, costs and harms due to removing human support may be passed onto customers. Cost-shifting of this nature unfairly burdens particular populations, such as people with disabilities or accessibility requirements. There is also cost-shifting (from AI service providers) to clients given intrinsic dependencies involved in adopting AI services. Apart from cost shifting, there are also labour implications.

The use of chatbot services may involve swapping call centre jobs with gig workers who train chatbot algorithms with computers purchased at their own expense and who work in their homes, subjecting their household to surveillance,[176] potentially with even less labour protection than a call centre worker.

The way in which AI enters the market of business operations is not often the focus of computer science or policy considerations.

Instead, computer scientists and policymakers typically consider how a chatbot interfaces with end-users (e.g. focusing on matters of data accuracy, safety etc.). But there is great value in examining the way chatbots -or other AI services - transform consumer relations, organisations and labour conditions, or redistribute risks to the weakest parties in its production cycle, such as gig workers doing menial tasks and end-users.

Indeed, how production is organised matters deeply and can impact certain groups or individuals over others. The same economic pressures on the business of computing may also affect the ability and willingness of tech companies to address their potential societal harms.

We expose in this section some of the issues that a production view, rather than an algorithmic or data-centric view, provides to policymakers. This section does not aim to be a complete insight into a production view of AI, as this remains understudied. This section intends to provide quick highlights of a framing that may reveal other ways in which AI has discriminatory effects or may intensify inequalities.

---

171 **Nick Kolakowski,** Google's Duplex Evolving: Can it meet your needs?, Dice, 8. May 2019. https://insights.dice.com/2019/05/08/googles-duplex-evolving-can-it-meet-your-needs

172 **Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, et al. 2018.** Applied machine learning at facebook: A datacenter infrastructure perspective. In 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 620–629.; **Arun Kumar, Matthias Boehm, and Jun Yang. 2017.** Data management in machine learning: Challenges, techniques, and systems. In Proceedings of the 2017 ACM International Conference on Management of Data. 1717–1722.

2.1 Dataset collection, data-ecosystem, and privacy

Dataset collection for machine learning is an expensive endeavour. For most machine learning, the data of concern is not just "personal data" but any data that is relevant for the operations that are going to be executed.

This includes data about operational resources, as well as data needed to optimise production, e.g. feedback on the use of computational resources. How the dataset is used after collection is also a source of cost, as we explain in the next subsection.

Organisations interested in AI must put great effort into reducing the cost of data collection and generation of labelled datasets. This has given rise to an industry which specialises in the production of datasets and a growing practice of using open datasets.[177]

Both approaches suffer from a number of problems, including data protection concerns, that we highlight below.

2.1.1 Data protection and privacy concerns

Due to market and cost-saving pressures, AI service providers can reduce costs by disregarding privacy concerns or data protection considerations when collecting data samples.[178]

Most companies will skirt privacy considerations by scraping public data from the Web, such as image datasets and text datasets.[179] This may not always be legal and, even if it is, might not be enough to fully address customers' normative expectations of privacy and meaningful consent.

Given that multiple data points can be combined from diverse sources, thereby revealing new information about individuals that was hidden from a primary dataset, AI services have the capacity to produce heretofore undiscovered privacy concerns.[180] Users usually give their consent for a specific context where they publish the data (often a social media), but they are not aware of the other potential uses.[181]

There are also more coercive scenarios in which public data are collected without user consent and then used to develop AI systems. Raji et al. also point out that the methods to collect samples might be dubious, giving the example of a start-up which signed an agreement with the government of Zimbabwe to collect face images from its population through various camera infrastructures, without the consent of the population itself.[182]

Policy that promotes debiasing, as well as the reduced availability of data for "minorities", may incentivise increased data collection of exactly those populations who may be vulnerable to surveillance.[183]

Given the above examples, the possibility that debiasing methods may lead to over-surveillance of marginalised populations should be a very serious concern.

Paradoxically the most recent European Commission proposal to regulate AI enables the use of sensitive attributes for debiasing, without further consideration of the risks it imposes on exactly the populations that the regulation says it intends to protect.[184]

2.1.2 Data about resources and operations

In addition to demonstrating the privacy implications of personal data collection, the production view also brings light to the public resource implications of data collection regarding systems.

When AI service providers collect datasets about operations, resources, geographic areas, and other systems that are essential to the management of public resources, there are knock-on effects for the control and management of public resources.[185]

AI service providers who move into this space risk disadvantaging or excluding groups and individuals who need access to such resources (e.g. to demonstrate and remedy distributive or other injustices, in order to get on with their daily affairs).

While the use of AI in logistical applications and resource management may increase efficiencies, it often comes with a trade-off in public control and oversight. When public datasets become the domain AI service providers, they acquire the power to allocate resources from a central vantage point.

The economic models of the business of computing also incentivise these actors to use their infrastructural advantage to tailor distribution of risks across populations for the sake of economic gain (see Section D.2.4 for how this may lead to forms of predatory inclusion and the intensification of economic inequalities).

While scholars have studied the market of data brokers and their impact on fundamental rights,[186] we have yet to see studies on how machine learning has given rise to companies and public initiatives focused on producing data for the optimisation and control of resources (e.g. satellite data for precision agriculture in Sub-Saharan Africa is organised by a number of players including Amazon and Microsoft),[187] with unexpected consequences for already disadvantaged populations.[188]

The ways in which these datasets privilege these parties in commodifying different aspects of life and resources, and how they may impact structural and global inequalities, must be central to understanding the impact of AI.

---

173 Our production view contrasts with the conception of AI as a knowledge infrastructure that subverts all power, as depicted in the recent book "Atlas of AI" by **Kate Crawford**. Aside from it not being a knowledge infrastructure, we argue that AI is a combination of technological and economic processes pushed by the business of computing dominated by Microsoft, Amazon, Google, Apple and Facebook and not the other way around. Moreover, the entanglement of the business of computing with "earth", "labour", "data", "classification", "states" and "space" predates the current wave of AI, and does not provide explanations for the computational and economic dominance that these few companies now exercise. Finally, by reducing all matters to "AI", Crawford introduces a critique that would be mute if the dominant computing paradigm (and marketing) pushed by industry would shift to another form, like it did from Big Data to Internet of Things before landing on AI, with quantum computing lurking in the horizon.

174 Here we are referring no longer to the Machine Learning View we presented in the previous section, but the family of computing techniques.

175 https://www2.deloitte.com/content/dam/insights/us/articles/4780_State-of-AI-in-the-enterprise/DI_State-of-AI-in-the-enterprise-2nd-ed.pdf

176 **Olivia Solon**, Big Tech call center workers face pressure to accept home surveillance, NBC News, https://www.nbcnews.com/tech/technews/big-tech-call-center-workers-face-pressure-accept-home-surveillance-n1276227

177 https://www.altexsoft.com/blog/datascience/best-public-machine-learning-datasets

178 **Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020.** POTs: protective optimization technologies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.

179 **Vinay Uday Prabhu and Abeba Birhane. 2020.** Large image datasets: A pyrrhic win for computer vision? https://arxiv.org/abs/2006.16923

180 **Jacob Metcalf and Kate Crawford. 2016.** Where are human subjects in big data research? The emerging ethics divide. Big Data & Society 3, 1 (2016), 2053951716650211. https://journals.sagepub.com/doi/abs/10.1177/2053951716650211

181 **Danah Boyd and Kate Crawford. 2012.** Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society 15, 5 (2012), 662–679. https://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878

2.2 Optimising machine learning pipeline costs

Machine learning pipelines go beyond obtaining datasets. They also require complex processes and computational environments for the efficient development, testing, and maintenance of the models and the systems they are part of.

For any organisation that is moving into machine learning, these are significant costs, often in the form of labour costs or capital expenses associated with computing machinery.

The industry has risen to this challenge by producing infrastructures for employing micro-workers (also discussed in the literature as gig-, task-, ghost- or crowd workers) at reduced costs, adopting cloud-based computational resources that shift capital expenses to operational expenses, deepening dependencies on mobile phone devices already rolled out to billions of users, and suppressing the production cost of computational hardware in the global supply chain.

We discuss the potential ways in which these production pipelines may cause inequalities or other harms below.

2.2.1 Labour conditions in production

Crowd-sourcing is employed in multiple activities of the machine learning pipelines, such as for data annotation, data filtering and, in some cases, even data collection.[189] Similarly, when users have troubles with workflows managed using machine learning, companies/organisations turn to low-paid micro-workers to make up for the failings of these systems.

Micro-workers, for example, are tasked with content moderation, technical support, customer relations, and responding to consumer contestations. Companies/organisations save money on labour costs in these low-end jobs in part by neglecting to care for workers or address harms to workers.

A study across 75 countries with 3500 workers found that despite micro-workers being necessary for the production of AI, "workers and their jobs remain invisible, poorly regulated and paid, seemingly not directly employed by the corporations that construct and run such systems".[190]

182 **Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020.** Saving face: Investigating the ethical concerns of facial recognition auditing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 145–151. https://arxiv.org/abs/2001.00964

183 https://www.theverge.com/2019/10/2/20896181/google-contractor-reportedly-targeted-homeless-peoplefor-pixel-4-facial-recognition

184 **European Commission. 2021.** 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. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

185 **Burke, M., & Lobell, D. B. (2017).** Satellite-based assessment of yield variation and its determinants in smallholder African systems. Proceedings of the National Academy of Sciences, 114(9), 2189-2194. https://www.pnas.org/content/114/9/2189

186 **Rieke, A., Yu, H., Robinson, D., & Van Hoboken, J. (2016).** Data brokers in an open society.

187 https://registry.opendata.aws/collab/deafrica; https://docs.microsoft.com/en-us/azure/industry/agriculture/overview-azure-farmbeats

188 **Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020.** POTs: protective optimization technologies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 177–188. https://arxiv.org/abs/1806.02711

Multiple concerns expressed by various human-computer interaction and social science literature revolve around the treatment of micro-workers within crowdsourcing platforms.

Crowdsourcing tasks are designed to prompt workers to be very fast at their job, accept a large number of tasks per day, pay a low payout, and depend on this job to make a living.[191]

This leads to exploitative behaviours by the annotation requesters [Paritosh et al. 2011]. Workers have low flexibility in time organisation,[192] they are automatically considered as unreliable if they refuse tasks, and their work is not valued.

Gig-workers not only sell their labour, but companies also require the exploitation of their personal assets (computer, car, bike, rented apartments, etc.).[193]

The data captured there is integrated as production data to increase the efficiency of the service providers operations. This data is also used to optimise worker productivity and labour costs, at times to the point of cruelty.[194]

Jamil and Noiseux argue that gig-workers, like uber drivers, are subject to a twofold process of "accumulation by dispossession".[195]

First, workers are dispossessed from labour protection, benefits and bargaining power.

Secondly, by giving service providers access to efficiently exploit their own assets (cars/phones/Internet connection), workers are "dispossessed from the value of their "dead labour" embodied in their private properties which are being monetized [Kenney and Zysman, 2016], exploited and consumed as part of the Uber process of value production." [Jamil, R., & Noiseux, Y. 2018].

These forms of dispossession are further amplified because gig-economies around the globe predominantly depend on racialised and migrant labour, an already structurally vulnerable population.[196]

While gig-work is a growing job sector, platforms that exploit this form of work have actively lobbied against increased worker protections.[197]

For example, recently approved Proposition 22, strongly pushed by Uber, exempts gig-workers from being considered as employees. Ongoing worker self-organising worldwide has contested this and sought to achieve fair conditions and designation as workers.[198]

When we consider that in many countries, gigworkers are predominantly from low-income and minority populations, it becomes clear how the production of AI sustains inequalities while dismantling social protections. To this day, platforms' accessibility for disabled workers, elderly people, etc. is low,[199] and the crowd workers' privacy is often at risk.[200]

The fair payout for crowd workers is a topic of research for crowd-sourcing platforms and has been supported by some of them recently, but not necessarily adopted by the requesters.[201]

Besides gig-workers, the working conditions of the teams that design and execute the pipelines, e.g. data scientists, data engineers, and MLOps engineers, and the political economic conditions of their labour are also of interest to the feasibility of applying debiasing,[202] but also the ways in which harms materialise in the design of AI.

2.2.2 Suppressing material costs and exploiting (natural) resources

Besides workers, the production pipelines reinforce the exploitation of resources, as these are fundamental to both developing and deploying them.

Most deployed systems require large amounts of data and computational power. These systems intensify reliance on fossil fuels.[203]

The ever-increasing need for energy for computing, amplifies existing inequalities and climate injustice.

Namely, politically, culturally and economically marginalised populations will suffer the consequences of climate change more severely. They will do so even though they use vastly less fossil fuel based energy, bear far less responsibility for creating environmental problems, and do not enjoy the benefits of technological innovations like AI to the extent that wealthier nations and people do.[204]

In addition to machine learning being computation-heavy, the collection of such data relies on the increased use of mobile devices and sensor networks, all of which require devices that depend on the mining of minerals and use of toxic materials.

Natural resources (as well as workers) around the globe are exploited in order to develop the hardware components, energy resources and infrastructures needed to build and deploy both the data engineering pipelines and the computational infrastructures for machine learning.[205]

The production of compute-heavy AI systems that depend on cloud and mobile computation reinforces environmental issues in areas of the world where resources are exploited, where data are hosted, and where computations are done.

189 **Prug, T.; Bilić, P.,** Work Now, Profit Later: AI Between Capital, Labour and Regulation // Augmented Exploitation: Artificial Intelligence, Automation, and Work / **Moore Phoebe V. ; Woodcok, Jamie (ur.)**. London, UK: Pluto Press, 2021. str. 30-40 https://www.bib.irb.hr/1116478

190 **Berg, J. et al. (2018).** Digital Labour Platforms and the Future of Work: Towards Decent Work in the Online World. At https://www.ilo.org/global/publications/books/WCMS_645337/lang--en/index.htm ; Irani, L. (2015a). Justice for 'Data Janitors'. At www.publicbooks.org/justice-for-data-janitors.; Prug, T.; Bilić, P., Work Now, Profit Later: AI Between Capital, Labour and Regulation // Augmented Exploitation: Artificial Intelligence, Automation, and Work / **Moore Phoebe V. ; Woodcok, Jamie (ur.).** London, UK: Pluto Press, 2021. str. 30-40.

191 **Joel Ross, Lilly Irani, M Six Silberman, Andrew Zaldivar, and Bill Tomachine learninginson. 2010.** Who are the crowdworkers? Shifting demographics in Mechanical Turk. In CHI'10 extended abstracts on Human factors in computing systems. 2863–2872. https://soundideas.pugetsound.edu/faculty_pubs/1009

192 **Ming Yin, Siddharth Suri, and Mary L Gray. 2018.** Running Out of Time: The Impact and Value of Flexibility in On-Demand Crowdwork. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 1–11.

193 **Jamil, R., & Noiseux, Y. (2018)**. Shake that moneymaker: insights from Montreal's Uber drivers. Revue Interventions Économiques. Papers in Political Economy, (60) https://journals.openedition.org/interventionseconomiques/4139

194 https://www.bbc.com/news/world-us-canada-56628745

195 **Harvey, D. 2004.** The 'new' imperialism: accumulation by dispossession. Socialist Register 40: 63-87. https://socialistregister.com/index.php/srv/article/view/5811

196 **Van Doorn, N., Ferrari, F., & Graham, M. (2020).** Migration and migrant labour in the gig economy: an intervention. Available at SSRN 3622589 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3622589

197 **Dubal, V. B. (2017).** The Drive to Precarity: A Political History of Work, Regulation, & Labor Advocacy in San Francisco's Taxi & Uber Economies. Berkeley Journal of Employment and Labor Law, 73-135. https://repository.uchastings.edu/faculty_scholarship/1589

198 See for example the work of the International Alliance of App-Based Transport Workers IAATW: https://iaatw.org

199 **Eunjin Seong and Seungjun Kim. 2020.** Designing a Crowdsourcing System for the Elderly: A Gamified Approach to Speech Collection. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 1–9.; **Kathryn Zyskowski, Meredith Ringel Morris, Jeffrey P Bigham, Mary L Gray, and Shaun K Kane. 2015.** Accessible crowdwork? Understanding the value in and challenge of microtask employment for people with disabilities. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing.

200 **Huichuan Xia, Yang Wang, Yun Huang, and Anuj Shah. 2017.** " Our Privacy Needs to be Protected at All Costs" Crowd Workers' Privacy Experiences on Amazon Mechanical Turk. Proceedings of the ACM on Human-Computer Interaction 1, CSCW (2017), 1–22.

The damage from these ranges from "water usage, pollution from backup generators, supply chains for the rare earth minerals used in hardware, and the toxic materials involved in the production of this hardware."[206]

The production, consumption and disposal of such resources require energy, increasing waste and pollution issues, while decreasing existing resources.[207]

2.2.3 Reducing engineering and management costs

Data collection, labour, and environmental factors are not the only contributors to AI use's (often unaddressed) costs.

The production view also draws attention to engineering and management costs and their relationship to social, political, and economic inequalities. In fact, the costs of AI rise due to how machine learning systems function (e.g. use of data in development and deployment) and are employed in practice (e.g. a centralised model for making predictions over many individuals, personalisation of model outputs through fine-tuning of a central model into many individual decentralised models, etc.).

Data and model engineering

Once data that is needed for the production of a machine learning model is identified, a 'data pipeline' is put in place to prepare the dataset for training and the evaluation of the machine learning model.

Data items might need to be labelled and filtered based on their relevance for the target inference task or based on material constraints (e.g. size of an image, blurriness), and processed (e.g. resizing of an image, modification of its colour scheme, etc.), both to respond to the technical constraints of the subsequent machine learning models, and to make the underlying algorithms train more accurate models.

In deployment, the new data for which inferences are to be made (such data are referred to as serving data, prediction data, inference data, or production data in literature) might also need to go through these processing steps, either simply for the model to be able to use them, or for maximizing their consistency with the training data. Similarly, exploring machine learning and building a model require numerous, costly steps.

Exploring the use of machine learning for a new application (experimenting with machine learning to find an economically viable model) requires trying out various datasets from various sources, processing this data in various ways, combining multiple datasets, as well as testing different machine learning algorithms and ways to train these.

Similar to this experimental phase, building a model follows a costly trial-and-error approach. Multiple underlying algorithms and sets of hyperparameters might be tested in order to achieve the best inference performance, various post-processing methods for the outputs of the models might also be tested, as well as multiple different ways to process the training data.

While there is no standardised process to develop and optimise a model, the amount of possible combinations of these design choices to ideally test is enormous, requiring intensive use of computation, with the issues discussed above.

Data and model management

As machine learning is used more and more to deliver services, optimising of compute resources becomes even more important.[208]

The compute and development costs from experimenting with or maintaining machine learning models can be immense.[209] Especially given the business risks involved in exploring machine learning applications, the promise of

cloud computing to shift capital expenses into operational expenses, by removing the need to invest in hardware and IT staff becomes ever more relevant for organi sations that want to deploy AI.

The cost and latency of data manipulations are so concerning, that by now infrastructures are put into place and become necessary for the use of AI.[210] Data storage and transfers from one engineering operation or model training operation to another (ingress and egress costs) are optimised within database management systems.

Data labelling is optimised through the creation and use of crowdsourcing platforms.[211]

We are now also starting to see the development of new companies, which offer specialised infrastructures for users of AI, that optimise data processing and allow for an easy integration of all data activities for machine learning purposes.

In the same way, the development and deployment of machine learning models are both becoming optimised through the rise of new infrastructures.[212] Development is for instance facilitated by simple, plug-and-play user interfaces, where computations are handled automatically in the back-end.

Deployment includes all activities beyond building an initial model in order to bring it to production and maintain it, i.e. model integration, monitoring, and revision (updates of the model when its performance starts to decrease).[213]

Deployment also requires putting in place infrastructures,[214] that allow for monitoring of the model performance and serving data, and updating the model, while using models that can scale to make inferences rapidly for large amounts of serving data.

New companies also propose the combined, automatic integration of both the data and model engineering pipelines, with either infrastructure that allows to efficiently support both, or software implementations that facilitate quick and easy development and deployment of systems.[215]

The engineering and management of data and models require costly investments even before identifying whether the resulting machine learning model will return on investments.

Even when successful AI applications are discovered, and the accounting tricks of the clouds are applied, some argue that due to the human support and material variable costs, the profit margins of AI can be low.[216] All of these cost factors and the complex production line they bring directly impact on the application of debiasing methods.

The cost of AI production is likely to either deter companies from catering to concerns about AI and discrimination, as this would require more computation, or reduce it to injecting a minimal debiasing activity into their pipelines for compliance purposes.

The complexity of these pipelines further raises serious concerns about the feasibility of effectively applying debiasing across all of these optimisation steps, a matter not yet considered in research.

Besides, due to cost constraints, drifts continuously arise from the data and model engineering pipelines. For instance, it is often the case that a model is trained on a readily-available dataset, or a dataset that has been collected through a simple, cost-efficient setup.

Yet, the new data in deployment are often captured in a very different way, leading to a data shift between the training and serving data, and potentially to issues around discrimination and unfairness.

For instance, in Detroit, following harmful mistakes of a facial recognition system in deployment, the police decided only to apply it to still images, as it is closer to the training data collected in a static setting in development.[217]

We also discussed data shifts and concept drifts in Chapter C.2. as they lead to questions about how these constant shifts impact debiasing and auditing? How often should these be performed? How costly would that be (especially in order to account for these drifts)?

---

201 **Nata M Barbosa and Monchu Chen. 2019.** Rehumanized crowdsourcing: a labeling framework addressing bias and ethics in machine learning. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–12.

202 **Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daum III, Miro Dudik, and Hanna Wallach. 2019.** Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. https://dl.acm.org/doi/10.1145/3290605.3300830

203 **Sarah Griffiths**, Why your internet use is not as clean as you think? BBB Smart Guide to Climate Change, 6. March 2020. https://www.bbc.com/future/article/20200305-why-your-internet-habits-are-not-as-clean-as-you-think

204 **Harlan, S. L., Pellow, D. N., Roberts, J. T., Bell, S. E., Holt, W. G., & Nagel, J. (2015).** Climate justice and inequality. Climate change and society: Sociological perspectives, 127-163. https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199356102.001.0001/acprof-9780199356102-chapter-5

205 **Thilo Hagendorff and Katharina Wezel. 2019.** 15 challenges for AI: or what AI (currently) can't do. AI & SOCIETY (2019), 1–11.; Aouragh, M., Gürses, S., Pritchard, H., & Snelting, F. (2020). The extractive infrastructures of contact tracing apps. Journal of Environmental Media, 1(2), 9-1. https://philpapers.org/rec/HAGCF-2

206 **Ingrid Burrington,** The Environmental Toll of a Netflix Binge, The Atlantic, 16. December 2015. https://www.theatlantic.com/technology/archive/2015/12/there-are-no-clean-clouds/420744

207 **Ben Williamson and Anna Hogan. 2020.** Commercialisation and Privatisation in/of Education in the Context of COVID-19. Tech. rep. Education International.; **Benedetta Brevini. 2020.** Black boxes, not green: Mythologizing artificial intelligence and omitting the environment. Big Data & Society 7, 2 (2020), 2053951720935141.; **Ravit Dotan and Smitha Milli. 2020.** Valueladen disciplinary shifts in machine learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 294–294. https://journals.sagepub.com/doi/full/10.1177/2053951720935141

208 https://builtin.com/artificial-intelligence/ai-computing-cost-reduction

209 https://www.altexsoft.com/blog/business/how-to-estimate-roi-and-costs-for-machine-learning-and-data-science-projects

210 https://www.eckerson.com/articles/data-management-best-practices-for-machine-learning; **Matthias Boehm; Arun Kumar; Jun Yang, , 2019.** Data Management in Machine Learning Systems , Morgan & Claypool

211 **Jennifer Wortman Vaughan, 2017.** Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research, 18(193):1-46, 2018. https://jmlr.org/papers/v18/17-234.html

212 **Polyzotis, Neoklis, et al. 2017.** Data management challenges in production machine learning.Proceedings of the 2017 ACM International Conference on Management of Data.; **Kumar, Arun, Matthias Boehm, and Jun Yang.** Data management in machine learning: Challenges, techniques, and systems. Proceedings of the 2017 ACM International Conference on Management of Data. 2017. https://dl.acm.org/doi/10.1145/3035918.3054782

213 https://exports.areviewz.com/wp-content/uploads/aws-managing-ml-projects.pdf

2.3 Externalities of optimising software production

Given how tech companies use machine learning to capture and optimise operations of other businesses or institutions (see Section above on public datasets, logistics, and operations), it should not come as a surprise that tech companies also use machine learning to optimise their own operations and meet their business goals.

This too has implications for inequalities throughout society, as optimised services create openings for mass manipulation, depend on live and potentially exploitative experimentation, privilege majoritarian behavioural patterns, and disappear minoritarian ones.

The rise of machine learning coincides with the data-centric production of software in the form of services. Unlike software that came out of a box and ran on users' devices, services and apps bind users into a long-term transaction with software companies a relationship constantly monitored and improved through user analytics.

Over the last two decades, machine learning has therefore become a fundamental part of software production, not only because of business models based on advertisements and user profiling, but because of the centrality of resource optimisation, AB testing and analytics to "disruptive" software production.[218]

Information systems today typically build on distributed service architectures and incorporate real-time feedback from both users and third-party service providers.

This feedback is leveraged for a variety of novel forms of optimisation that are geared towards the generation of value through the system. Often times, optimisation of resources is part and parcel of end-user-facing functionality, e.g., autocomplete in search can increase user satisfaction and improve query processing performance including the reduction of expensive memory calls.

Machine learning has also become part and parcel of "continuous development" strategies based on experimentation that allow developers to define dynamic objective functions and build adaptive systems. Businesses can now design for "ideal" interactions and environments by optimising feature selection, behavioural outcomes, and resource planning in line with a business growth strategy.

For example, social networking sites like Facebook continuously refine software features, like tagging in photos; to optimise user engagement.

They can use machine learning to set up large-scale experiments to select the design of the tagging feature that brings more users back onto the platform, e.g., one design could inform users that they have been tagged but require them to come back on the platform to untag themselves.

While for the users, a tagging mechanism may give them a feeling of control over their interactions on the platform, the greater visitor numbers this feature enables, especially if users then also spend time on the platform, maintains the ideal conditions for their ad delivery business.

Machine learning-based software systems developed using such strategies are increasingly developed to capture and manipulate behaviour and environments in order to generate value.[219]

Turbo powered by AI, software designers can experiment with and iterate on designs that capture optimal populations and activities that can be tied to value generation. Such an ordering of populations and activities for the purpose of value generation leads to the "sorting individuals based on their estimated value or worth".[220]

Researchers found a more recent example of this in the way Facebook optimises ad targeting which results in women on the platform being more expensive to advertise to due to higher Click Through Rates (CTRs).[221]

It would be mistaken to interpret this as a system that promotes women or values them more. Rather, such a system evaluates the value of Facebook profiles based on their advantage, or exploitability, for their ad operations.

Debiasing methods intend to address panoptic sort partially: they promise to achieve fairness in optimisation, while keeping intact "the commodification of everyday activities", for example, of interactions on online social networks.[222]

Moreover, these systems may introduce broader risks and harms for users and environments beyond the outcome of a single algorithm within that system.

While the layers of optimisation introduce efficiency and allow systems to scale, they also pose social risks and harms such as mass manipulation, majority dominance, minority erasure[223] and media addiction.[224] These systems may fulfil some notion of fairness, and may not even harm any individual significantly, but may cause harms at scale.

Finally, an AI service can be optimised to be fair to its users but introduce harms to environments and people beyond the system.

For example, location-based services like Waze provide optimal driving routes that put users in certain locations at a disadvantage.

Waze often redirects users off major highways through suburban neighbourhoods that cannot sustain heavy traffic.

While useful for drivers, it can increase overall congestion, or affect neighbourhoods by making streets busy, noisy, and less safe. Consequently, towns may need to manage, fix and police roads more often.[225]

---

214 https://www.netapp.com/pdf.html?item=/media/16114-ar-idc-infrastructure-considerations-for-aipdf.pdf

215 https://neptune.ai/blog/end-to-end-mlops-platforms

216 https://a16z.com/2020/02/16/the-new-business-of-ai-and-how-its-different-from-traditional-software

217 https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong-how-facial-recognition-led-to-afalse-arrest-in-michig?t=1616003785339

218 **Seda Gurses and Joris Van Hoboken. [n.d.].** Privacy after the agile turn. ([n. d.]).

219 https://www.altexsoft.com/blog/business/how-to-estimate-roi-and-costs-for-machine-learning-and-data-science-projects

220 **Gandy Jr, O. H. (2021).** The panoptic sort: A political economy of personal information. Oxford University Press.

221 **Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019).** Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-30. https://arxiv.org/abs/1904.02095

222 **Gandy Jr, O. H. (2021).** The panoptic sort: A political economy of personal information. Oxford University Press.

223 **Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March).** On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). https://dl.acm.org/doi/10.1145/3442188.3445922

224 https://www.theguardian.com/news/series/cambridge-analytica-files

225 **Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020.** POTs: protective optimization technologies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 177–188.

2.4 Exclusion, predatory inclusion and AI

The optimised mode of production that is intrinsic to machine learning is an important determinant of what services are produced and for who. The outcomes of these political, economic choices can produce results that counter typical liberal notions of inclusion as a desirable outcome.

Digital inclusion strategies have been on the agenda of policymakers that hope to address inequalities through the introduction of internet-based services, that promise but fall short of providing a global network. Similarly, the high cost of machine learning production, as well as situations where machine learning favours larger datasets and hegemonic knowledge bases, may come to intensify some forms of exclusion from future services, reinforcing global inequalities.[226]

These forms of exclusion raise concerns around equal access to AI services within the EU, a matter that is definitely not addressed by debiasing.

However, having access to or being included in machine learning systems may not always be desirable. Many community activists from Black and migrant communities have, for example, powerfully raised this point with respect to facial recognition.[227]

Machine learning may also create a tier system in who has privileged access to services. Machine learning is promoted to reduce costs by replacing costly operations with services scaled up using machine learning. Such systems can create tiered systems where some people have access to services with human experts, and others only have access to ML-driven solutions propped up by gig-workers.

One example of such inequality can be seen in the health sector. In a proposal for "flipping the Stack" on health care, the authors argue that the health sector could be more cost-efficient and innovative, if instead of giving people direct access to care delivery, an operation that can be supported by ICT, we could develop systems that would invert that experience.[228]

In the flipped stack, users first use a health app tied to sensors that monitor their health and, if needed, connects them online to health professionals, minimising hospital visits.

The disputable efficiency gains and safety concerns aside, one can imagine such systems producing societies in which live access to care professionals is the luxury of a few, and a health care profession increasingly indistinguishable from other gig workers.[229]

Finally, the efficiency gain of optimising resources and behavioural outcomes, if applied at scale in exploitative systems, can lead to predatory forms of inclusion.[230]

The South African governments' dependency on Cash Paymaste Services, owned by the technology firm Net1, for the digital distribution of welfare payments can be regarded as a recent example of such predatory inclusion.[231]

The company used its data advantage, profiling capabilities and monopoly power to create a private marketplace in where Net1's other

subsidiaries could market products and services, making deductions directly from welfare recipients' welfare accounts.[232]

Facebook and many blockchain companies' quest for reaching the unbanked through digital platforms, international organisations and corporations cooperating on biometrics and payment systems for refugees, or industry interest in UBI bring the power of computation to financial systems, with the potential to scale predatory financial practices to some of the most vunerable populations in the interest of profit.

2.5 Policy implications

Despite the extensive literature on machine learning as a method or technique (e.g., for classification, prediction etc), and tech companies as business (e.g., focusing on the market activities and business models of these companies), there is little detailed analysis of its intersection with the production of such systems.

A greater understanding of the business of computing could provide better explanations why we experience certain phenomena, (e.g., why is this industry pushing for AI?), and whether policies or methods to address the downsides of this business are adequate (e.g., is debiasing sufficient to address the harms of AI across its production pipelines?).

Above we showed the potential impact of the economic pressure on the production of AI-based systems. The growth mandate that tech companies are subject to, given their rather high valuations, pushes forth computation heavy methods like machine learning and Blockchain.

However, while it promises organisations cost cuts, the production of AI-systems is costly and complex.

226 https://cordis.europa.eu/article/id/421665-equity-in-natural-language-processing

227 https://edri.org/our-work/reclaim-your-face-biometric-surveillance; https://www.wired.com/story/defending-black-lives-means-banning-facial-recognition; https://www.ajl.org; https://racialjusticenetwork.co.uk/our-work/stop-the-scan-campaign

228 https://thehealthcareblog.com/blog/2020/04/11/flipping-the-stack-can-new-technology-drive-health-cares-future

229 https://www.theguardian.com/commentisfree/2020/oct/28/england-coronavirus-covid-test-and-trace-teenagers

230 **Taylor, K. Y. (2019).** Race for profit: How banks and the real estate industry undermined black homeownership. UNC Press Books.

231 **Foley, R., and M. Swilling.** "How one word can change the game: case study of state capture and the South African Social Security Agency (p. 82)." Stellenbosch: State Capacity Research Project (2018).

232 **Taylor, L. (2021).** Public actors without public values: legitimacy, domination and the regulation of the technology sector. Philosophy & technology, 1-26.

Typically, this creates dependencies on big tech companies that offer computational infrastructures and services that can be used in the production of machine learning (e.g., Tensorflow, AWS, Amazon Mechanical Turk).

It is by now also common to find companies that try to successfully address and optimise the cost and complexity of different parts of a machine learning pipeline (e.g., focusing on labeling or building specialized infrastructures).

While these smaller companies may aim to alleviate technical complexity, they increase the complexity of governing what can be considered machine learning supply chains. The number of parties involved in the production pipelines, the lack of well-defined processes, and the low margins of AI projects, all raise great challenges to the feasibility and effective application of debiasing methods.

Given the production costs, the way companies apply debiasing may turn out like privacy-by-design: debiasing frameworks may be picked up, sometimes only performatively, when they add to company's bottom line or increase computational dependencies.

Regulating AI production and use

Despite dominant narratives about automation and the promise of cost reductions, the production of machine learning is labour and compute-intensive.

Making machine learning a reality passes through exploitative labour practices and extractive supply chains, while increasing our dependencies on fossil fuels.

The production of AI is therefore already thoroughly entrenched in global inequalities and climate injustice, a non-issue in most policy documents. It is notable how little production harms are considered in AI policy making, or by the industry players that promote de-biasing methods.

The production view may help bridge advocates and activists working on the seemingly disparate but deeply connected topics of labour, migration, extraction, climate, and AI.

The production view also hints at the ways in which AI may further concentrate the business of computing in the hands of a few tech companies. Many organisations will be pushed to adopt AI but will not be able to afford to develop their own systems. Already the business literature advises organisations to take off the shelf AI-based services, or to reuse models.

However, when they do so, it is unclear if these organisations will be able to apply debiasing metrics that are specific to their contexts and populations. If not, we may see organisations picking up AI services with guarantees that a "generic" debiasing metric has been applied for a universal population that is not theirs.

3. Other viewpoints on AI

In this section, we outline two other viewpoints relevant to policymakers, advocates and activists.

> **When applied in the context of public institutions like education, health, or transportation, the integration of machine learning ties organisational missions to AI services providers' and Big Tech companies' mandates to grow or return on investment. Under these conditions, public institutions become both success of technology companies: a co-dependency on unequal terms.**

3.1 Infrastructural view

When talking about the infrastructural view, we are making reference to the assemblage of institutions and processes that makes the machine learning industry thrive and, ultimately, dominate.

This viewpoint necessarily asks us to consider the ways in which AI systems drive an ever-expanding model of growth, privileges large players, and orients the AI industry towards consolidation and highly centralised control.

To be clear, the lens of computational infrastructure forces us to consider the connection between computational infrastructure and "plain old" internet infrastructure and, more importantly, what drives agenda-setting in the AI industry. Indeed, the production and deployment of machine learning are heavily dependent on existing and developing computational infrastructures, e.g., mobile phones running iOS or Android, cloud infrastructures, and sensor networks, all of which are dominated by a few companies.

These infrastructures are the result of a tech industry with great market valuations, which puts these same companies under pressure to continuously grow. Indeed, machine learning is an important part of these companies' growth strategies through their computational infrastructures.

Machine learning requires compute-heavy applications, including the networked connectivity that such applications demand. The more widespread it becomes, the more machine learning intensifies the demand for data collection and for computation.

Furthermore, ownership of and control over key parts of computational infrastructures give these (now legacy Big Tech) companies privileged access to production data, as well as an advantage in shaping machine learning practices.

An infrastructural view exposes the growing power asymmetries people and institutions face vis-a-vis large tech companies that have access to and make available vast stores of data, computing resources, and machine learning capabilities. Large companies centralise power by virtue of dominating access to data, storage space, computational power to process them, financial resources to afford the resources needed for developing machine learning pipelines.[233]

Some scholars highlight the general power inequalities that develop from the barriers to enter the field of machine learning as the required resources are unaffordable to most individuals or small organisations and companies, and possibly governments.[234]

More recent articles argue that computational infrastructures "make it possible for companies to keep growing while lowering their fixed costs structure" removing barriers to market entry to companies.[235]

Moreover, we find that research and innovation in machine learning are also becoming concentrated in the hands of these same key players, since their computational resources far outpace what any independent institution can provide. In short, Big Tech firms are dictating rules of the game and creating new and diverse path dependencies in terms of market development and governance of technology production.

Unfortunately, discussion and response to these infrastructural dependencies are virtually absent from most policy and research analyses.

The infrastructural view helps elucidate the scale of risks and harms once key players consolidate control and ownership over computational resources.

Furthermore harms from the use of AI in systems can be multiplied when datasets and models become infrastructural, i.e. when the same data set or model is used by many applications or many parties due to the cost of production. Once sunk into infrastructures, categories and orderings carried out with machine learning are harder to contest or remove, as more parties come to depend on them.

The machine learning APIs made available by large companies, freely or in return for a fee, necessarily reflect a narrow set of bias objectives selected according to market demands, with little opportunity for their users to change these or even to be aware of the contextual implications of infrastructural decisions.

In such scenarios, machine learning service providers that secure infrastructural positions then become arbiters of political contestations, as evident for example in matters of content moderation on social media platforms.

How far these providers may respond to these contestations may depend on a cost function that does not always align with just outputs or the public interest.

The integration of ML services into computational infrastructures deeply impacts the ability of public and social actors, and especially members of marginalized communities, to question these infrastructures or demand red-lines, prohibitions or other governance responses mitigating the harms of AI systems.

3.2 Organisational view

This last view attends to the more precise dependencies and asymmetries felt by organisations as they become entangled with AI systems and focuses on the ways in which AI applications complicate organisations' ability to serve the public interest or provide the conditions for the exercise of fundamental rights.

Many organisations view AI as a "revolutionary" way to automate and augment organisational workflows and operations. Packaged with the promise of scalability, efficiency and effective problem solving, AI-based systems offer organisations the possibility to automate and centralise workflows, and optimise institutional management and operations.

Due to the political, economic conditions highlighted in the production and infrastructural views, these transformations are likely to bring about dependencies on third-parties and computational infrastructures, including on their economic models and licensing schemes.

An organisational perspective helps elucidate how AI-driven third-party services are procured (or sometimes simply introduced when members adopt their services in organisational workflows), implemented and deployed in a climate of pragmatism and instrumentalisation.

233 **Ravit Dotan and Smitha Milli. 2020.** Value-laden disciplinary shifts in machine learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 294–294. https://arxiv.org/abs/1912.01172

234 **Stevie Chancellor, Shion Guha, Jofish Kaye, Jen King, Niloufar Salehi, Sarita Schoenebeck, and Elizabeth Stowell. 2019.** The Relationships between Data, Power, and Justice in CSCW Research. In Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing. 102– 105.; **Zeynep Tufekci. 2015.** Algorithms in our midst: Information, power and choice when software is everywhere. In Proceedings of the 18th acm conference on computer supported cooperative work & social computing. 1918–1918.

235 https://medium.com/aperture-hub/strategy-in-the-post-fixed-costs-economy-fe2caab957f8

In practice, the dependency on third-party services and computational infrastructures potentially places the autonomy of these organisations at risk in visible and less visible ways.

When applied in the context of public institutions like education, health, or transportation, the integration of machine learning ties organisational missions to AI service providers' and Big Tech companies' mandates to grow or return on investment. Under these conditions, public institutions become both dependent on and instrumental to the economic success of technology companies: a co-dependency on unequal terms.

In general, the distribution of costs of failure and success are profoundly uneven for tech companies, public institutions, and citizens, not only due to cost-shifting in risks and harms to the end-user or organisations [see Section D.2.3 above], but also because such a dependency upends the democratic safeguards that form the raison-d'etre of public institutions.

Beyond issues of financial dependency, the adoption of AI by public institutions cuts right into the execution of operations and the ability of these institutions to serve the public.

For example, with the pandemic, as universities come to depend on third-party services for their remote learning and administrative needs, they not only stop investing into their own technical infrastructure in favor of those provided by the third-party service providers, but they also commit to swapping their organisational IT and online education know-how for license managers.[236]

This puts universities on a dependency path for more third-party services as they continue to digitalize, as they will eventually have undone their institutional know-how to implement local solutions. These shifts expose public education institutions to the unbundling and rebundling of their fundamental operations by market players.[237]

Depending on third parties for educational services also changes the makeup of public education, foregrounding individualised pursuit of mastery enacted primarily through AI, in favor of education that, for example, promotes interpersonal dialogue and relations with others.[238]

How these developments will play out for the democratization of education, closing of educational disparities, and the future of public education, are questions part and parcel of these seemingly technical decisions.

If successful, AI may deliver organisational transformations designed to divert much wealth from public and private institutions will pass onto technology companies.

How strong these organisations will remain once their operations and workflows have become dependent on the computational infrastructures, and the companies that run them, is still to be seen. What impact this may have on inequalities in our societies, however, is not a wait-and-see question, but one that requires great attention and advocacy.

---

236 **Fiebig, Tobias, Seda Gürses, Carlos H. Gañán, Erna Kotkamp, Fernando Kuipers, Martina Lindorfer, Menghua Prisse, and Taritha Sari. 2021.** Heads in the Clouds: Measuring the Implications of Universities Migrating to Public Clouds. https://arxiv.org/abs/2104.09462

237 https://a16z.com/2020/10/16/next-gen-edtech

238 **Ben Williamson and Anna Hogan. 2020.** Commercialisation and Privatisation in/of Education in the Context of COVID-19. Tech. rep. Education International. https://issuu.com/educationinternational/docs/2020_eiresearch_gr_commercialisation_privatisation

E.

Conclusion and recommendations for civil society and policymakers

1. Summary

The policy documents we reviewed recognise that the use of AI can cause grave harm.

In Chapter B, we sampled soft-policy documents from EU institutions that make gestures to the potential impact of AI on discrimination of protected groups and the widening of social inequalities. That these issues made it so prominently into policy documents suggests that politicians and public institutions recognize that these issues are serious and inseparable from the introduction of AI.

What is surprising is that the same policymakers who recognise the gravity of the problem have landed on debias -ing (data) as the way to mitigate these structural concerns.

To provide a first intuition of the mismatch between the dimensions of the problem and its potential solutions, in the rest of Chapter B, we include a summary of select policy recommendations regarding AI and inequalities. We couple this with an overview of state of the art in research on debiasing, specifically outlining the limits of such techniques to these ends.

We conclude that policy documents lack genuine engagement with existing theories, activism or laws around structural discrimination.

'Discrimination', 'equal access' and 'structural inequalities' are used interchangeably and are not grounded in existing EU law, social theory, or informed by current social movements.

They further fail to cover the basics and the limitations of the science of debiasing. The documents lean towards debiasing of datasets as the ultimate goal, often demonstrating a lack of consideration for biases that may occur in algorithms and their outputs.

Generally, the documents assume that debiasing can be applied universally, when the budding field of research has only touched a small set of social domains, often through a US-centric conception of discrimination and inequalities. They implicitly pursue the creation of "unbiased datasets" as the ultimate solution to AI and discrimination.

While mitigating against biases in statistical terms is the goal of technocentric approaches, the technical infeasibility, social undesirability and flattening of different political positions inherent to the pursuit of "unbiased datasets" have escaped regulators.

Overall, as we showed in chapter B, policymakers do not provide sufficient guidance on debiasing requirements, and how to address their

technocentric limitations. They also treat AI systems like a packaged product, pushing the complexities of AI production pipelines and the continuously evolving services they deliver out of regulatory scope. In sum, it is difficult to assess the validity of the current policy demands, or the effectiveness of its future application.

In Chapter C, we show the narrowness of debiasing approaches and discuss the state of the art in debiasing research.

Debiasing approaches oversimplify complex problems of injustice, or even politics. Even if these simplifications can be insightful for scientific inquiries, debiasing can be difficult to apply and insufficient as the sole basis of audits to capture the discriminatory effect of AI-based systems.

The bias frame is potentially useful as a post-hoc tool to identify those discriminatory effects that can be detected. However, it should be recognised that while this is a necessary and limited inquiry, applying these methods is far from ensuring that the algorithms or systems into which they are integrated are "free from discriminatory effects".

By design, debiasing approaches privilege service providers to implement debiasing solutions at their own discretion. Paradoxically, computer scientists have developed frameworks that assume those parties economically incentivized to align the application of AI with their bottom line will also take the necessary steps to provide a remedy for the discriminatory effects of their products and conduct audits.

As such, debiasing frameworks privilege service providers in deciding what counts as discrimination, when it occurs, and when it is sufficiently addressed.

Finally, we show that by focusing on data and algorithms, most debiasing approaches limit their focus to inputs and outputs of these systems rather than on their impact in the world, letting service providers off the hook with respect to the broader impact of these systems on discrimination and inequalities.

For all of these reasons, policymakers should at best consider debiasing as a minimum effort to flag blatant technical issues, but not consider "bias" as the only problem, or debiasing as the main solution, to AI's discriminatory impact and its broader consequences for societal inequalities.

While it is beyond the scope of this document, the lighttouch approach to debiasing in European policy documents, the limitations of debiasing methods, as well as the regulatory implications of the proposed technical solutions, require further attention. Debiasing approaches, especially in the US, are aligned with deregulatory approaches to privacy; which argue that collection of data is inevitable, and instead, regulation should focus on use and the accountability of service providers.[239]

The European policy documents could also be seen as indicative of a permissive regulatory environment to come, in which service providers are likely to "twist process to serve corporate ends".[240]

Under these conditions, the assumption that addressing the discriminatory effects of AI can and should go through service providers is especially concerning.

While the researchers that develop these frameworks may have positive objectives, in practice, debiasing approaches coming out of the research are likely to reinforce the power of service providers to decide what are acceptable societal harms, while also weakening regulatory protections and enforcement (e.g., by normalizing the collection of (sensitive) data, or by certifying that systems are bias-free) when it comes to the harms of AI.

If debiasing approaches fall short of addressing the way in which AI, and the tech companies that will benefit most from their deployment, may come to reorder societies unfairly, what then?

To provide a better grasp of the technical, economic and political impact of AI, Chapter D highlights different ways in which AI-based systems may come to reconfigure our societies in harmful ways. In particular, we take a closer look at the potential issues associated with machine learning as a technique, and the production of AI based systems.

We also provide the basic contours of framings that expose the political economy of computational infrastructures dominated by a few tech companies and what it means to bring AI into organisations.

Due to the production costs, AI-based systems will be deployed by or dependent on the computational infrastructures of big tech companies.

When applied in the context of public institutions like education, health, or transportation, integrating machine learning puts computational infrastructures and their economic mandate to grow, or return on investment, into the heart of institutions that are tasked with serving the general public.

This creates a series of dynamics, including creating a direct connection between public institutions and the economic success of technology companies.

This is, however, not just a financial dependency, but, through the adoption of AI, one that cuts right into the execution of their operations and their ability to serve the public. The impact of AI-based systems on the governance, operations and financial stability of these organisations are likely to challenge the ability of our societal institutions to provide individuals with the necessary conditions to exercise their fundamental rights.

Going beyond its impact on individuals, a political, economic view that accounts for the infrastructural and organisational shift that AI-based systems bring about may help identify a broader set of legal and policy mechanisms to regulate AI and tech companies.

---

239 From Collection to Use in Privacy Regulation? A Forward Looking Comparison of European and U.S. Frameworks for Personal Data Processing, **In: Van Der Sloot, Broeders and Schrijvers (eds.)**, Exploring the Boundaries of Big Data, Netherlands Scientific Council for Government Policy, 2016, pp. 231-259

240 **Waldman, A. E. 2019.** Power, process, and automated decision-making. Fordham L. Rev., 88, 613. https://ir.lawnet.fordham.edu/flr/vol88/iss2/9

2. Gaps in policy-making

In this report, we investigated and outlined debiasing as researched and practiced in computer science, and we demonstrated the limited engagement of these documents in matters of discrimination and debiasing. This led us to identify a series of issues and recommendations for policymakers, advocacy groups and other civil society actors.

> **The impact of AI-based systems on the governance, operations and financial stability of organisations are immense, and could bring about transformations that challenge the ability of our societal institutions to provide individuals with the necessary conditions to exercise their fundamental rights.**

2.1 Problems with debiasing as a policy response to structural discrimination

The field of fairness which focuses on developing and evaluating debiasing methods is still in the making. For the purposes of this report, it is important to remember that in computer science:

"Bias" refers to the ways in which a dataset or the outputs of a machine learning model can be skewed. A discriminatory bias (or unfairness) more specifically refers to the lack of parity in these data or outputs for different population groups or for different individuals in the population.

The study of biases in ADMs and debiasing methods have been for a small set of often US-centric usecases, and a small set of inference tasks, such as inferring whether someone who committed a crime in the past is likely to reoffend.

These examples do not necessarily generalise to other domains or to other regions of the world and cannot be considered universal solutions that can be applied independently of context.

Debiasing works by first choosing a bias/fairness metric on which the system should perform well, and then by applying a debiasing algorithm to the system.

The bias metrics and debiasing methods, as well as the level of difficulty to apply them, differ depending on the type of data employed (e.g. text,

tabular, images) and on the type of inference task the system performs (e.g. classification, regression, recommendation, ranking, set selection). This impacts the viability of debiasing solutions even for the simplified problem of parity they are trying to address.

There currently exist three main sets of bias metrics, with more than 20 metrics in total. These are: group bias metrics (comparing outputs on the basis of two groups characterised by a protected attribute); individual similarity metrics (measuring the extent to which similar individuals even within the same group are treated similarly) and causal metrics (the causal relations between the protected attribute(s) and the model outputs in the inference process). There are trade-offs between these metrics that also have social consequences.

Based on state of the art in machine learning, debiasing and AI policy-making, we conclude the following:

Conclusion 1

Policy documents and the biased framing in computer science are not aligned. Current policy documents are imprecise when discussing the problem in terms of discrimination, and when turning to bias mitigation and auditing as solutions. Across the board, policy approaches do not account for the gap between the two

Unclear notion of discrimination renders it difficult to qualify the problem

Policy documents do not identify specific types of discrimination, and use the terms 'discrimination', 'equal access' and 'structural inequalities' interchangeably without clearly grounding them in existing EU law, social theory, or demands of current social movements. In contrast, bias metrics focus on particular occurrences of discrimination, and are limited in complexity by the ways these occurrences can be quantified.

The ambiguity in terminology and lack of clarity on how these may translate into computer science concepts makes it hard to evaluate the extent to which debiasing methods answer the problems that the policy documents target.

The lack of conceptual differentiation further creates uncertainty in how to respond when AI systems have discriminatory effects that do not have a simplified technical counterpart in debiasing, e.g., intersectionality.

Narrow focus on debiasing data and models leaves out harms from machine learning production activities

Policy documents primarily refer to datasets, and sometimes machine learning models, as potentially biased entities.

They rarely mention underlying algorithms or other activities in the production pipeline of AI systems, e.g. data labelling and processing activities. In doing so, the documents implicitly make the assumption that discrimination solely occurs in datasets or models, and that mitigating their respective biases at one point in time allows building non-discriminatory systems based on them later on.

These assumptions have been shown, in research and in practice, not to be correct especially due to the complex interactions between components, and due to the various feedback loops of AI systems, once again leaving the complex choice of debiasing methods unguided and other bias issues that arise in the lifecycle of AI systems out of regulatory scope.

Debiasing is still a science in the making, with many limitations not recognised by policymakers

Debiasing methods generally consist of modifying one of the three main components of an AI system in order to make the system's outputs closer to the selected metric: the data on which a machine learning model is trained; the objective

function that serves to train the model; or, the way its outputs are processed. Each of these methods sees its effectiveness limited, and can be challenging to apply in practice.

Besides, there is no clear guidance in order to select the most appropriate method for each usecase. Policymakers however neither recognise these limitations nor provide clear guidelines to apply the methods, leaving open-ended what application of debiasing methods may count as an effective mitigation of biases.

In summary, policymakers' engagement with the discriminatory impact of AI falls short socially and technically. The documents fail to demonstrate a clear understanding of discrimination, do not capture most existing debiasing methods, do not recognise the trade-offs of these methods, nor do they provide guidance or the necessary conditions for independent and meaningful audits that may complement their application.

Debiasing is a rich field of research that is in the making. It is favoured by an industry incentivized to solve complex socio-technical problems they introduce by using more of their own tools.

This burgeoning field is, however, far from having convincing results, straightforward applications, or holistic frameworks that could be considered "solutions" to the discriminatory effects of AI based systems.

The field of research, its practitioners, and policymakers would benefit immensely from decentering technocentric framings of the AI discrimination problem, and should aim for a more sincere engagement with AI and its reordering of society.

Conclusion 2

Policy documents confer wide discretion to technology developers and service providers to set the terms of debiasing practices, leaving out challenging political and economic questions of these methods to the discretion of service providers

Socio-technical considerations necessary for the application of debiasing are left out from technocentric policies

Policy documents present debiasing, a family of methods that contain many social assumptions, as simple tools to apply.

This gives service providers free reign over the interpretation of the socio-technical considerations that should be taken into account before their application (e.g. around the choices of metrics and their implications, around the entity having the responsibility to make these choices, such as developers, model requesters, public institutions, etc.).

While some of this imprecision may be intended to make policies technology-neutral, it also delegates sensitive decisions to technology developers, highlights primarily data debiasing as a sufficient target, while suggesting the complexities of debiasing approaches may be solved within technical standardisation bodies.

Neither discrimination, nor the complexities of debiasing can be tackled only by technology specialists, solely based on data quality considerations. A sincere response from policymakers should require experts from other disciplines to account for the complexity of discrimination, centre-affected parties and should go beyond algorithms to consider a more holistic evaluation of AI systems.

The choice of debiasing metrics may reveal results that can favour different actors of the system, be it individuals or groups impacted by the system, the service providers or public interest

If we imagine a bank making loan decisions, applying debiasing metrics to false negatives over false positives, might reduce the risk of the

bank to the detriment of the clients of the bank. In contrast, disparate impact might better match the expectations of society for fairness at the expense of model accuracy, a factor that may impact the profits of the bank.

Metrics are not only dependent on the social context, but also on other qualities of the data and the model, a matter that neither policymakers nor commercial debiasing tools provide guidance on. This lack of guidance is likely to give rise to strategic application and auditing of debiasing methods by service providers.

Policymakers are vague on matters of auditing and transparency, and do not define or guarantee the conditions necessary for independent audits

The policy documents under-specify what needs to be audited, when, and by who, reinforcing the difficulties in applying debiasing methods narrowly.

Policymakers should be more concrete in their requirements towards audits, must guarantee the conditions for independent audits, demand audits to go beyond technocentric evaluations. It should also be clear that these audits cannot be considered sufficient to guarantee that AI based systems are free of discriminatory effects.

Multiple (commercial) debiasing tools embed and make invisible the limitations and political, economic underpinnings of debiasing

A number of academic initiatives and companies have developed tools to support the application of bias metrics and debiasing methods. These tools serve to make debiasing techniques more accessible to analysts, but also embed their limitations, assumptions and political and economic incentives into the development of "fair" systems.

The promotion of these tools as market - able solutions erases the limitations of the debiasing approaches and is likely to result in the application of the approaches with unpredictable results. Ideally, such tools should be subject to greater scrutiny by researchers and regulators alike.

Most concerning, existing policies put the service provider in the driving seat in matters of discrimination and inequalities

By shifting the solution to complex socio-technical problems of discrimination into the domain of design, and by promoting debiasing frameworks that give the discretion to address these to service providers, policymakers deem technology companies arbiters of societal conflicts.

Such policy-making is likely to strengthen the regulatory power of tech companies in matters of discrimination and inequalities, while normalizing the application of AI-based population management methods aligned with profit objectives.

Conclusion 3: Policy documents do not recognise the conceptual and practical limitations of debiasing and bias auditing

Conceptual limitations

The bias framing limitations in its conception of discrimination as it focuses solely on parity in models outputs in relation to protected attributes. This is often based on information already available in training datasets.

In the debiasing literature, most examples pertain to ADMs that make inferences about people (inference subjects).

Further bias frameworks take a statistical rather than an individualized view on how populations of inferences subjects are considered in the system.

The focus on the ideal output distribution to be achieved, leaves out considerations as to how individual inference subjects may be harmed otherwise by the model.

The focus on inference subjects in the inputs and outputs of the algorithms means that people and elements in the environment of the system, that are not inference subjects, but may be impacted by the system, are not considered in the analysis of harms. Discrimination based on intent, or structural inequalities, are also left out of scope.

To summarise, we recognise especially the following conceptual limitations of debiasing:

Narrow scope of CS research compared to the breadth and depth of AI discrimination

Research on fairness in CS has tackled very specific problems in select domains, often based on applications of ADMs in the US. This means that the knowledge base of debiasing research is specific to these applications and may not generalise sufficiently to be used as solutions for many other domains, in other countries, in which discrimination may be a concern.

Moreover, AI is applied beyond ADMs in applications that may have unexpected discriminatory effects.

Inference subjects are considered members of broader, simplistically defined groups with parity ideals

In terms of the individuals the bias metrics focus on, their ideal outputs are not accurately reflected by the metrics. The metrics only support the desire for parity. Such parity can only be controlled between two groups, and complex problems, like intersectionality, are handled in a simplistic manner. Yet, depending on the situation, individuals within a group might be affected differently by the inferences.

This information however cannot be understood from the sole use of attributes and labels, as it is about the impact of a label on each individual separately, and not on the present characteristics of the individual.

The metrics that are precisely dedicated to account for such issues (individual fairness metrics) are also limited in that if they would precisely account for structural discrimination, they would also need to categorize individuals based on potential discriminatory disadvantages they might be subject to, leading to see these individuals once again as members of larger groups.

All these metrics also ignore individual justice where individuals should be considered individually and not in comparison to others – what machine learning instead does.

Lack of consideration of actions performed based on system outputs despite their impact on inference subjects

The bias framing focuses on the outputs of the algorithms, and does not necessarily reflect potential impact of humans (or secondary systems) in the loop on the inference subjects – which may not be captured in the data.

Hence, bias auditing or mitigation may miss the discriminatory or other harmful effects due to the combined outcome of humans or secondary systems in the loop making use of algorithmic outputs.

Lack of debiasing methods that target structural discrimination instead of sole, observable, outcomes

Debiasing does not necessarily solve the structural causes of the discrimination that might have happened before the introduction of the AI system, and which can be perpetuated at scale even with a debiased system due to the sole focus on outcome distributions and protected attributes.

Practical limitations

Existing debiasing and bias auditing methods are challenging to apply in practice, due to the difficulty of gathering relevant data, and due to the complex, interdependent organisation of the components of AI systems and of their development process.

Lack of access to data necessary to apply debiasing or bias auditing

The type of information required to apply debiasing or audits (e.g., sensitive attributes, predictions, training data) might be difficult to access.

This is typically due to the organisation of the AI production pipeline, lack of access to the production environment, or trade secrets. If independent auditors do not have easy access to this data, and at times to the production environment or pipeline, audits are going to be very malleable by service providers.

Contentious requirement to access sensitive data for bias auditing or debiasing

The sensitive information required to apply debiasing (e.g., sensitive attributes, predictions, training data) might have harmful downstream consequences. Especially after the holocaust, there has been great vigilance in Europe concerning the use of race and other sensitive attributes in administrative systems.[241]

The use of machine learning for administrating populations, be it by public or private organisations, rekindles this contentious issue and touches on an already existing conflict between administrative organisations in EU and activist communities.

Activists in Europe and elsewhere have been demanding from public and private institutions to keep statistics based on different identity attributes, e.g., in policing, in order to be able to provide evidence of discriminatory practices.

However, this demand for measuring discrimination should not be equated with the building of administrative AI systems based on race, gender or other sensitive attributes.

Moreover, the categories used for these measurements are deeply political and require much greater discussion before normalizing the use of sensitive attributes by AI providers, as evidenced in the recently proposed AI regulation.

The integration of sensitive attributes into the production of systems or into audits should be thought through very carefully and applied with great caution than is currently the case. Any decisions on these topics should also be made in light of the fact that debiasing methods are likely to intensify the surveillance of exactly those populations that these approaches are claimed to help protect.

Risk of missing issues to audit for

It is impossible, especially for model developers, who often do not have the domain knowledge sufficient to understand a model application's use-cases, to anticipate all harms that could arise from a system prior to its deployment.

This is where the constraints of the knowledge base of debiasing currently limited to select domains, mostly based on case studies in the US become apparent.

Applying debiasing in European contexts will require much greater political, historical and domain knowledge than is currently available.

The absence of this knowledge is not coincidental, but also a consequence of European knowledge traditions, marginalization of work on topics like race, ethnicity, migration, disability, and the demographic segregation prevalent in research and policy institutions.

Many-hands problem in the production of AI

The way the development and deployment of AI systems are organised in production pipelines and supply chains makes it hard to identify the parties that should be responsible or held accountable for debasing or auditing AI components.

Lack of knowledge in research on the handling of real, deployed systems

The deployment process of AI systems with many components and feedback loops is more complex than the simplistic models considered in research settings. Neither scientific research, nor policymakers, provide insight or guidance on debiasing in deployed systems (e.g., how to handle debiasing in transfer learning or model revisions).

Limited efficiency of bias mitigation methods

Due to the typical, statistical limitations of machine learning, and to the existence of inherent trade-offs between different fairness and accuracy metrics, bias mitigation methods cannot lead to achieving entirely unbiased models. Yet, such technical limitations are also not considered in policy documents.

Downstream trade-offs between metrics, their assumptions and the stakeholders they advantage

Machine learning systems are subject to unavoidable trade-offs that prevent service providers from verifying multiple bias objectives or bias for more than two groups at a time. This requires service providers to make trade-offs between metrics and between populations.

Currently, no mechanisms exist to ensure those trade-offs are done fairly and accountably. Further, we lack guidance on when such trade-offs are unacceptable, or short of a just outcome, when AI-based systems should not be deployed.

Beyond debiasing

Even if practitioners apply debiasing methods as demonstrated in carefully crafted academic papers, these methods contain conceptual and practical limitations. In an attempt to "represent" complex social phenomena in the design of AI-driven systems, these methods push conceptions of society that flatten social complexities and, in the process, entrench classification of populations based on hegemonic identity markers. By focusing on inputs and outputs of algorithms, debiasing frameworks suggest they can separate the impact of AI systems from existing structural inequalities.

Research in the field takes little note of how these systems may redistribute risks, cause harm to their environment and intensify inequalities in the world, even when producing "fair outputs".

By design, debiasing methods empower the service provider to determine what counts as debiasing in AI systems. This may seem like a win vis-a-vis a service provider that could otherwise just optimise for pure accuracy.

However, this also gives service providers the power to determine how much debiasing is optimal for their business. Using debiasing frameworks, a service provider can easily audit its system and return an "unbiased" result by carefully crafting the dataset and metrics used for the audit.

Given these limitations, is there still a role for debiasing and audit in policy-making? Yes, but with caution, and depending on future developments in research, regulation and society. Neither debiasing nor audits can certify the lack of harm, or even guarantee the reduction of harm. However, they can be used as part of a broader and more robust auditing framework for discovering systems that may pose discrimination risks.

For that, policymakers should not only pay lip service to these methods, but support the creation of the appropriate legal and technical conditions to perform meaningful system audits that can be conducted independently.

Looking forward: Alternativeviews on tackling AI and its potential impact on social, economic and political inequalities

To go beyond the narrow framing of AI, its discriminatory impact and debiasing as a solution, Chapter D introduced various viewpoints on AI-based systems.

We called these viewpoints because they look at the same phenomenon but from different conceptual vantage points, revealing different ways in which AI-based systems have already been reordering societies, with the effect of causing discrimination and intensifying inequalities, at scale.

Of the four viewpoints, we discuss the Machine Learning and Production views on AI in greater detail.

Machine learning view

Current policy focus on ADMs misses that the application of AI beyond ADMs may also have discriminatory effects and intensify inequalities. By scrutinizing the fundamental principles of machine learning applications, we surface the potentially harmful assumptions made when adopting machine learning more generally.

Policy documents and the biased framing put aside issues introduced by the use of machine learning, in terms of data collection and privacy, of dubious optimisation tasks, data attributes and label taxonomies, and of implicit repetition of data patterns.

Especially, a model might have been debiased, but the task that it serves to accomplish might not be sound or desirable, resulting in more harms (e.g. fairly allocating bad working conditions to job seekers). Besides, machine learning consists of learning and repeating patterns in past behaviours, leading to the exclusion of previously unknown populations, and to privilege decisions by comparing individuals to other individuals captured in datasets–which is not always desirable.

Machine learning also requires to categorising things and people, and finding (proxy) data to represent them and the inferences to make, often leading to harmful oversimplifications in this representation exercise.[242]

The use of pseudo-science, e.g., eugenics, phrenology, physiognomy, in task modeling, the choice of (proxy) attributes, e.g. the Fitzpatrick scale to denote race, and the selection of populations in Machine Learning models further raise red flags. The bias lens does not account for any of these potential issues.

Production view

This viewpoint shows how looking at the business of computing, rather than computer science as a scientific field, can provide a deeper understanding of the societal changes that the use of AI is likely to bring about.

In particular, we look at the way in which machine learning promises to cut costs and optimise day-to-day operations of an organisation, solving complex resource allocation or logistical problems, or improving production lines in many sectors ranging from manufacturing to creative industries.

This means that many applications of AI will take place in Business to Business (B2B) contexts, and not just in consumer-facing (B2C) applications. We use this viewpoint to demonstrate how the intense use of machine learning to allocate resources, or to turn complex social activities into resource allocation problems, e.g., delivery of jobs, news, pizza, or dates, is likely to have great impact on social ordering, experiences of discrimination and inequalities.

The production of AI typically requires collecting and introducing new flows of data, a costly process which has given rise to a cottage industry of data processors, with implications for privacy, security and resource allocation.

We bring to the foreground how making machine learning a profitable reality intensifies the exploitative labour practices and extractive supply chains that underlie the tech industry. These practices paradoxically include concentrating the vulnerable people that debiasing frameworks pretend to protect in mines, factories and in a rapidly growing gig-work sector, under harmful working conditions.

Even when there is a genuine benefit to the user of these services, the data engineering and management costs are likely to bear negatively on the application of debiasing methods, as well as the feasibility of applying meaningful audits on these systems.

These examples also bring into view how the optimisation of software features at scale help companies to capture and manipulate people and their environments, with harms beyond discrimination that come to being when such systems are applied at scale, e.g., election manipulation, misinformation, targeted advertisement of subprime loans, or traffic congestion.

Finally, we discuss how these systems can flip liberal notions of inclusion and exclusion by making it possible to implement business models that benefit from greater inclusion of economically or otherwise marginalized populations in exploitative systems. e.g., systems that aim to integrate the unbanked.

The production view builds in political economy into the machine learning view. Once machine learning is put to use in the business of computing, accuracy metrics are redefined to measure the efficiency and profitability of business operations, rather than the accuracy of representations.

For instance, with emotion detection or facial recognition, what matters in reality is an ability to "improve the efficiency of operations", e.g., for the operation of a targeted advertisement infrastructure or a policing establishment respectively.

From a business perspective, the validity of the task is not necessarily dependent on the scientific basis of the task, or the accurate representation of emotions or people, but on the efficiency of the operations. Debiasing approaches offer a way to reintroduce representational concerns AI based systems may raise while optimising towards the value generation objectives of the entity deploying the system.

As a result, while they may improve the representational outcomes, debiasing approaches leave the operational priorities, as well as the political and economic consequences of these systems unquestioned.

Further strong incorporation of broader viewpoints is required

Finally, we sketch two additional views that highlight the role of two important players: the providers that dominate our current computational infrastructures that AI applications depend on to enter our daily lives, and organisations that are in the process of adopting AI, e.g., universities, hospitals, as they are likely to inform further points of policy intervention. A more in-depth study of these viewpoints is needed. In the infrastructural view, we start exploring ways in which AI may further concentrate the business of computing spearheaded by big tech.

Over the last decade, the computational infrastructures concentrated in the hands of a few companies have promised to become the technical and financial engine of the business of computing.

Ownership and control over these computational resources give these companies privileged access to data, as well as the ability to shape machine learning practices. Moreover, compute-heavy AI applications help to tighten societal dependencies on their computational infrastructures.

The material implementation and the production costs of AI applications are likely to reinforce these dependencies. As is currently the case in domains already dominated by big tech, the concentration of machine learning models in the hands of big players is likely to make them arbiters of political contestations and social justice questions in domains like health, education, policing.

All of this points to how seemingly technical dependencies may come to strengthen the political, social and economic inequalities due to the accumulation of infrastructural power in the hands of big tech companies.

If AI as a project succeeds in maintaining this dependency, it could potentially lead to a remarkable transfer of wealth and political power to tech companies in the coming years.

In the organisational view, we discuss how organisations, and not just individual users adopting AI services, are likely to drive the adoption of AI. While AI may bring benefits to organisations, it also gives rise to organisational challenges due to automation, commodification, economic models, and the use of techniques for operational control.

We briefly explain how structurally and economically, the introduction of AI-based services, and the dependency on external services and computational infrastructures they bring about, potentially place at risk organisational and economic autonomy of organisations.

When applied in the public domain, like in education, health, or transportation, integrating machine learning puts computational infrastructures and their economic mandate to grow, into the heart of the institutions that are tasked with serving the general public.

This creates a series of dynamics, including creating a direct connection between public institutions and the economic success of technology companies.

This is, however, not just a financial dependency, but through the adoption of AI, one that cuts right into the execution of operations and the ability of these institutions to serve the public.

The impact of AI-based systems on the governance, operations and financial stability of organisations is immense, and could bring about transformations that challenge the ability of our societal institutions to provide individuals with the necessary conditions to exercise their fundamental rights.

---

241 **Farkas, Lilla.** "Analysis and comparative review of equality data collection practices in the European Union: Data collection in the field of ethnicity." Directorate-General for Justice and Consumers Directorate D–Equality Unit JUST D 1 (2017). https://ec.europa.eu/newsroom/just/redirection/document/45791

242 see https://sites.google.com/view/beyond-fairness-cv/home for an example of computer science workshop that attempts to go beyond debiasing and surfaces some of these issues.

3. Recommendations for policymakers

In summary, EU policy documents on AI show that to date policymakers have failed to genuinely engage with the structural discrimination brought by AI-based systems as well as the science of debiasing.

By narrowly focusing on the technocentric solution of debiasing algorithms and datasets, and not recognising its limitations, this narrow approach squeezes complex socio-technical problems into the domain of design and thus into the hands of technology companies.

This approach empowers service providers as arbiters of discrimination and inequity, a paradoxical proposition. Overall, current AI policy-making in the EU underestimates the inequalities that may materialize with AI and the way its application reinforces computational infrastructures in the hands of Big Tech.

In light of these shortcomings in AI policy -making, as well as other viewpoints presented above, we make six recommendations for policymakers, researchers, advocates and activists, and propose adopting broader frameworks that look beyond data and algorithms in engaging technology companies.

By narrowly focusing on the technocentric solution of debiasing algorithms and datasets, and not recognising its limitations, this narrow approach squeezes complex socio-technical problems into the domain of design and thus into the hands of technology companies.

3.1 Policymakers adopting technocentric approaches to address the discriminatory impact of AI must define problems clearly, set criteria for solutions, develop guidance on known limitations, and support further interdisciplinary research

3.1.1 Policymakers should engage with and learn from prior work on eliminating discrimination and inequalities as part of identifying problems to tackle

Policy-making in this area would benefit from a deeper understanding of structural social, economic, and political inequalities in Europe and elsewhere, and of past successful regulatory interventions.

Such an engagement is likely to provide a better grasp of the problems that AI may bring, and a more accurate assessment of whether technocentric solutions are sufficient to address the complexity of these problems.

3.1.2 Policymakers should better acquaint themselves with the basics and limitations of debiasing approaches before proposing them as solutions in regulatory interventions

Policymakers must go beyond a datacentric understanding of bias and debiasing.

Debiasing applies to models and their outputs as well as to datasets. At the same time, policymakers must improve upon their understanding of debiasing as a solution to all discrimination: debiasing is a narrow technique that applies to a limited set of machine learning technologies in order to optimise for simple conceptualisations of bias.

3.1.3 Policymakers should provide clearer guidance on applying debiasing and independent audits

This guidance should recognise the limitations of debiasing and bias auditing methods, and include technologies and companies as well as additional regulatory mechanisms to mitigate the limitations of technocentric approaches.

Policymakers and researchers should work together to make the state of the art and the limitations of bias auditing and debiasing more accessible. Since many researchers work for or receive funding from big tech companies, measures to avoid conflicts of interest should be applied to such collaborations.

3.1.4 Policymakers should demand that any evaluation for discriminatory impact couples analysis of bias in an AI systems outcomes with an assesment of overall system objectives

Debiasing literature artificially separates outcomes and system objectives. In other words, it is possible

to have harmful systems that give fair outputs. For example, we could have systems that fairly distribute harmful jobs, as those at Amazon Warehouses, or allocate subprime credit.

Amazon Warehouse Jobs as well as subprime credits have been primarily targeted at people from financially and otherwise vulnerable populations, mostly people of color. Parity in such systems of exploitation neither makes sense nor is it a desirable state. Evaluations should therefore include both the system's objective and its potential for harming certain populations.

3.1.5 Policymakers should support interdisciplinary research on holistic approaches to auditing AI systems for discriminatory effects

Auditing the complete supply chain over time is likely to raise challenges that current research and auditing practices do not address.

Research should develop holistic auditing frameworks that address the challenges of deployed AI systems and should build a set of guidance tools to support practitioners in applying these frameworks.

The development of this research should also take into account the abundant research on intersectionality, anti-Blackness, etc. Researchers should aim at developing such frameworks for applications and domains that have not yet received much interest despite the harms they may create.

Next, policymakers should also support research on how to implement internal auditing and monitoring, including studies of the advantages and pitfalls of such technical approaches.

As auditing for discrimination in systems is a socio-technical process, auditing frameworks should be designed to involve technical developers, domain experts and system stakeholders.

3.2 AI policies must limit the discretion of AI service providers in addressing discrimination and inequalities

3.2.1 Policymakers should support an effective, decentralised system of assessing AI systems, discrimination and inequalities

Leaving bias auditing in the hands of service providers makes it technically hard to validate, and limits it to capturing simplistic statistical definitions of harms.

This approach also further empowers service providers as arbiters of discrimination and inequity, a paradoxical proposition.

Given the potential impact of AI on all aspects of society, we need actors, technical tools and observation techniques that assess AI independently of powerful public and private institutions.

These actors need to bring together expertise and stakeholders working on structural discrimination and societal inequalities, financial structures and procurement rules, and they need to be given the enforcement power necessary to evaluate, prevent, contest and mitigate the harms of AI systems.

3.2.2 Policymakers should refocus the bias attention onto bias audits

While it is not sufficient by itself, for now evaluating bias in AI audits remains a necessary part of the policymaker's toolset for controlling damage.

However, instead of considering bias audits of data, models or outputs as a sufficient goal, policymakers should consider using bias audits as part of a more comprehensive audit for identifying harmful systems that should be subject to further scrutiny and limitations.

In computer science, the framing of bias can be used as a pre-filter for investigating rather obvious harms before digging into more profound issues. This can be done during design and deployment, and in context, relying on ex-ante and ex-post audits that evaluate the potential harms of the system, given its objective(s) and implementation.

3.2.3 Policymakers should ensure that audits can be conducted independently

Policy measures must urgently create the appropriate conditions for independent audits (e.g., access to data and production pipelines), provide relevant guidance or criteria for the effective use of techniques to evaluate bias while auditing, and augment such policies with measures that counter the limitations of technocentric auditing for bias.

The results of these independent audits should be underpinned with enforcement mechanisms that require correction or constrain or ban systems that fail their audits.

3.2.4 Policymakers should set hard limits on access to sensitive data for auditing or debiasing

There is a danger that the process of applying debiasing methods may lead to AI-based (administrative) systems that optimise resource allocation or deliver services based on (hegemonic) attributes used to classify populations.

Neither the development of AI-based systems more generally nor debiasing specifically should become an excuse to allow service providers to collect sensitive data or to design systems using it.

Further, we need mechanisms of oversight when sensitive attributes are used for independent audits. Finally, more research on audits is needed, that is not limited to finding technical biases, that is not constrained by the need to have sensitive attributes, and that is cognizant of having potentially erroneous data.

3.2.5 Policymakers should avoid increasing surveillance of minorities or vulnerable populations in the name of debiasing

In the interests of achieving fairer results, debiasing may require collecting more data from populations that are underrepresented in datasets.

This means subjecting those populations to greater surveillance and exposing them to greater risks from powerful public and private institutions .

Solutions to this problem should not be solely based on technocentric utility calculations, but based on a principled approach to privacy and the needs of affected communities.

3.3 AI regulation needs to go beyond ADMS, data and algorithms to include the spectrum of AI applications and the broader harms associated with the production and deployment of these systems

3.3.1 Policymakers should expand the evidentiary scope of harms to non-technical criteria

Advocates should insist that evaluations of AI-derived inequalities should include both technical and non-technical assessments.

These should take into consideration the desirability of the machine learning task, complexity of the context of its application and the organisational and financial tensions that accrue due to AI optimisation, and do so in a way that centres the experiences and needs of those who are most impacted by the introduction of AI.

3.3.2 Policymakers should expand the scope of who (or what) may be classified as an affected party or AI subject and how they are harmed

Machine learning affects more than the "inference subjects" (subjects whose data is used to train the machine learning system or people about whom the system makes inferences); harms may extend to people who are not the direct subjects of an AI system's outputs.

For example, a system may harm individuals by reorganising aspects of their lives, or limiting their access to resources. Further harms can accrue to workers and to people's environments due to the way AI production lines are organised.

Individuals may also suffer from environmental harm caused by the large resources many AI systems require.

Finally, they may be harmed by the structural power dependencies and financial relationships that AI systems reinforce. All these harms must be made explicit, addressed in regulatory frameworks, and included in independent audits.

3.3.3 Policymakers should address distributed harms, exclusions and predatory inclusion through AI-based systems

AI systems have been used to produce effects at scale that harm whole communities even when they do not verifiably harm individual persons.

Examples include AI-based manipulation of elections, social sorting that bars economically undesirable populations from accessing AI-based services, and the inclusion of vulnerable populations in exploitative financial systems. Large-scale harms require policy attention and regulatory mechanisms.

3.3.4 Policymakers should ensure that auditing extends across the supply chain of AI production and captures the evolution of services

One-time audits assume AI systems are products. In reality, AI is a process developed through a production pipeline whose supply chain comprises many parties and that may incorporate more than one AI system.

Similarly, AI-based services are often provided by multiple actors and undergo continuous evolution that cannot be captured by a single snapshot. Further, data or domain shifts may create many harms in the course of AI's production or deployment.

Researchers should support policy-making by developing processes that address the challenges of auditing AI systems in production and deployment.

3.3.5 Policymakers should require that AI services available through application programming interfaces (APIs) are audited by service providers in the contexts in which they are deployed

APIs may be used by dozens or hundreds of organisations, or millions of people, in different contexts, (e.g., Amazon's cloudbased Rekognition computer vision platform is used by hundreds of government agencies).

Applying an audit only to the output of the APIs of such widely-used services does not capture the resulting biases that may arise in different contexts.

Governance measures such as public impact assessments should be developed to evaluate the structural impact of introducing these AI systems and should include an assessment of the validity of the system, and its contextual outcomes (e.g., an organisation may use an AI-system for its own purposes, or to serve a population, and may use the API for unforeseen purposes or for populations vastly different from those used to train the AI model the API delivers. All these contexts need to be considered).

3.3.6 Policymakers should bring harms accrued in the production of AI into the scope of regulations

The production of AI-based services includes further harms (e.g., labour conditions in production, the concentration of low-income/minority workers in gig-work, the environmental damage of extractive industries).

Broader regulatory frameworks are needed to evaluate and respond to these harms. Such considerations should feed into decisions about whether and how AI systems are deployed, and should form part of procurement decisions.

3.3.7 Policymakers should ban the deployment of AI services that reproduce biological essentialisms and fascist, racist or supremacist conceptions of humans and societies

There are many precedents for the use of essentialising or supremacist assumptions in machine learning. For example, the use of debunked or discredited science (e.g., eugenics, phrenology, physiognomy) in task modelling has led to claims that AI systems can infer sexuality or criminality from images, and also to the use of proxy attributes, such as the use of the Fitzpatrick skin type scale, to denote the highly sensitive category of race.

The proposal to use debiasing to address the outputs of these systems underlies how the limitations of debiasing may lead to absurd outputs, reinforcing harmful systems. The development and deployment of such systems are highly contested and should be banned.

3.4 AI policies should empower individuals, communities and organisations to contest AI-based systems and to demand redress

3.4.1 Policymakers should enable the contestation and banning of harmful AI-based services

Policymakers should implement regulatory processes so that AI systems which are inherently harmful or contrary to the public interest can be limited, prohibited or halted whilst in use.

They should create supervisory organisations that can support communities affected by the (illegal) deployment of such systems in contesting their deployment and receiving redress.

3.4.2 Policymakers should enable affected parties to trigger internal and independent audits

Organisations adopting AI-based systems and the people subject to their outputs should be empowered to trigger internal and independent audits.

The system's reevaluation should be made publicly available and processes should be available to enable the contestation of both systems and results.

3.4.3 Policymakers should ensure that audits of AI systems include and empower affected parties

Similar to debiasing, auditing AI requires a good understanding of a system and its context.

This is only possible if domain experts, organisations affected by AI-driven transformations, end users and affected communities, are involved in the process (while avoiding 'predatory inclusion' through participatory debiasing methods).[243]

243 We borrow the term predatory inclusion from Keeanga Yamahtta Taylor who uses it in the US context in which abolishing 'redlining' led to further discrimination and entrenchment of racial inequalities in the housing market.

3.5 AI regulation cannot be divorced from the power of big tech companies to control computational infrastructures

Addressing the rise of this infrastructural power requires long-term strategy and planning.

3.5.1 Policymakers should include within AI policy the broader impacts of the introduction of AI through computational infrastructures

The extractive impact of AI-based systems in terms of labour and natural resources, the organisational shifts on deployment and infrastructural dependencies AI enables, and the rise of Big Tech are all connected.

These factors must be balanced against the 'benefits of AI' when policy priorities are set, particularly those relating to investment, funding and regulatory priorities.

Policy priorities designed to encourage beneficial AI must not overlook the question of infrastructure dominance and weaken regulatory attempts. Instead, policymakers should pursue a broad harm prevention approach, and develop an innovation policy that provides alternatives to dependence on Big Tech.

3.5.2 Policymakers should invest in research on the production of computational infrastructures and the political economy of Big Tech

Research on the social, economic and political impacts of AI systems is unlikely to be funded or promoted by industry players.

Besides developing further methods for auditing AI systems that have not been studied until now (e.g., applications of AI in agriculture or transportation), we urgently need interdisciplinary research into the production of AI and how the financial and infrastructural developments of tech companies impact organisations, societal institutions and different communities.

3.6 AI regulation should protect, empower and hold accountable organisations and public institutions as they adopt AI-based systems

3.6.1 Policymakers should grant rights of redress to organisations that deploy or are affected by third-party AI services and depend on computational infrastructures

Organisations that adopt AI services should be empowered to demand customised machine learning models for their contextual need.

Organisations should be held accountable for contextual (discriminatory or harmful) outcomes of AI systems regardless of whether they develop these themselves or procure them from third parties.

Given the power asymmetries, some protection and rights of redress should be provided to organisations when they encounter conflicts with or sustain harms from providers of AI services or computational infrastructures.

3.6.2 Policymakers should assess and build the capacity of public and private sector organisations to deploy AI while mitigating its broader harms and inequalities

Understanding the greater structural impact that will result as AI enables the tech industry to move into other sectors requires better understanding of the current state of affairs in those domains.

This includes building capacity to evaluate the necessity and integration of AI in a way that improves these organisations' ability to serve people and their environments, as well as their ability to address inequalities.

Many institutions do not have the capacity to evaluate the broader impact of AI on organisations; this matter calls for urgent capacity building.

4. Reflections for advocates and activists

We would like to reserve final words for advocates and activists. Throughout this report, we show that biases based on data and algorithms and ADMs do not address the greater inequalities that may occur with the introduction of AI systems.

This has great consequences for advocacy and tech activism. While debiasing work has helped raise popular consciousness about the inequalities inherent to AI systems, the framing of the harms of AI as one of bias has also limited the space for action.

As demonstrated in all the efforts on algorithmic design, e.g., projects on debiasing, explanation, and transparency, the algorithmic view has limited many efforts to reformist action that sees solutions to problems of technology in developing more (slightly less harmful) technology, ultimately with the same companies.

This approach has also obfuscated the harms inherent in the production of AI that is all around us. We hope, especially the political, economic shifts we outline in the viewpoints, can open other spaces for technology critique and engagement, as well as rethinking our theories of change.

While we aimed to de-center technology, even the way we set up the alternative viewpoints circled us deeper into the technology and the companies that produce them.

However, rather than looking at the impact of AI on individual users, human rights, or communities, we introduced a light sketch of an organisational view as our attempt to explore other ways of decentering technology.

Given current conditions, calling to task organisations to adopt technology in way that serves people and addresses inequalities, seemed like a more productive option than the "community participation in AI" mantra that has become a common utterance in certain circles. This is our way of pushing back on frameworks promoted by tech companies that erase the role of existing organisations, and undermine their societal responsibilities, as if AI service providers are the only entities that serve communities.

To put this in an even greater context, the term AI also refers to a multi-trillion dollar technology investment that is currently moving financial markets.

With such great numbers as its burden, the business of computing continuously pushes companies and governments to invest resources and attention onto advancing the application of AI.

This funneling of resources into AI impacts the immediate availability of funds for other public needs. If successful, AI will channel more funds to tech companies. Regardless of its market success, AI will do so at the cost of greater justice questions.

Looking forward, COVID-19 and the recent economic stimulus packages are likely to push public organisations to digitalise. AI is going to be one important way tech companies are going to try to capture some of this spending.

This raises a number of vital questions: How can we ensure that digitalisation of institutions does not mean that they become a pass-through for already-dominant tech companies to capture more populations and greater access to the management of vital resources?

How can we ensure that this potential collaboration between public organisations and tech companies does not lead the prior to reduce complex social matters they are tasked with to those that can be executed by AI systems of (operational) control?

Most importantly, how can we ensure that the stimulus funds, and the digitalisation that may come with it, are used to ensure public institutions are strengthened to address existing inequalities and serve social-justice-oriented results?

We hope these questions will not only de-center technology, but also bring together communities that are already sharing computational infrastructures, to coordinate on other possible futures.

This funneling of resources into AI impacts the immediate availability of funds for other public needs. If successful, AI will channel more funds to tech companies. Regardless of its market success, AI will do so at the cost of greater justice questions.

Apendix

A. Prelude: Basic machine learning concepts

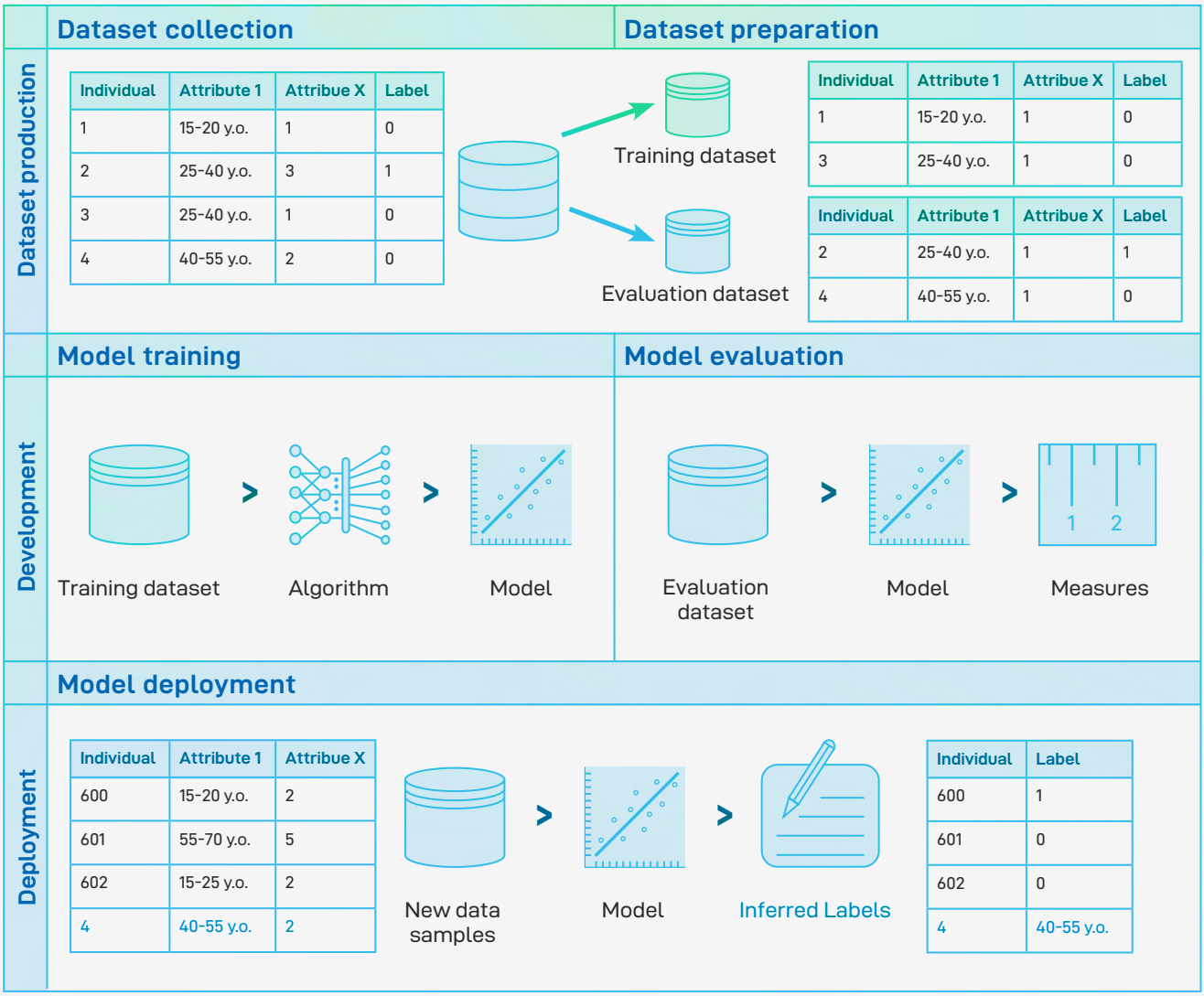
We deem it necessary to introduce the reader to the basic principles around the development and deployment of machine learning techniques in this section, to be able to better understand the discussions around bias in computer science in section 3.

Feel free to skip this section if you already have an idea of the general functioning of machine learning. We focus on this part of the report on machine learning because this is where most scientific works have been showing interest.

A.1. The machine learning formal setup

Developing and deploying a machine learning model is typically done in three to four steps, summarised in Figure 3.

Figure 3: The machine learning formal setup.



A.1.1 Dataset production

First, a dataset is produced, constituted of example samples and the labels that should be associated with it.

For instance, for facial recognition, the samples would be images of faces of various individuals, and the labels would be their names if their identity is targeted.

To do so, data samples are collected and in necessary cases annotated with labels, and later processed depending on the constraints or criteria of use for the machine learning application (e.g. size of images, removal of potential missing values or outliers in the data, etc.).

The dataset is divided into a training set and a test set (there can also be a validation set).

A.1.2 System development

The data samples are all encoded into a vector representation (also called a set of features) that a machine learning algorithm can support.

The encoding process is either performed by manually engineering features (e.g. computing the relative position and size of the eyes, nose and cheekbones for facial recognition), or by automatically learning features from a dataset (e.g. often the raw image pixels are inputted and a deep learning model is used to transform the pixels into a more informative description of the image samples for facial recognition).

A machine learning algorithm is selected and trained with the training set, its hyperparameters being fine-tuned using the validation set. The trained algorithm forms a machine learning model.

The model is evaluated using the test set, by checking how many of the samples are correctly labeled by the model.

There can be variations in the exact metrics used for this evaluation phase, but they usually correspond to some measure of accuracy, i.e. the percentage of data samples for which the model associates a correct label.

A.1.3 System deployment

Later, the model is deployed for performing its task in its real context. The model has inputted previously unseen data samples on which a label is expected to be inferred, and it outputs the inferred label.

A.2 Machine learning metrics

Understanding bias metrics requires understanding concepts around the evaluation of machine learning models.

We introduce these concepts here, as they are necessary to introduce for the reader to grasp the meaning of bias metrics and debiasing methods, and to identify limitations in their interpretations.

In Figure 4, we summarise these concepts and illustrate them through an example.

Let us introduce these concepts based on one typical example in bias and fairness literature for machine learning that we discussed earlier, recidivism prediction.

> **Example: Recidivism prediction - System functioning.** In this use-case, an entity wants to know whether an individual who committed a crime previously is likely to recidivate or not. For that, the entity collected data samples corresponding to the descriptions of various individuals, and ground-truth labels, i.e. labels attributed to each individual indicating whether this individual indeed re-offended (positive label) or not. It then built a machine learning model trained on this dataset, and tasked to infer the risk level of new individuals as a label.
>
> When the model is applied to individuals, it can infer risk labels that are correct or wrong. Depending on the individual on which the inference is made and on the correctness of the label, the inferences are coined differently.

Figure 4: The machine learning typical metrics, computed from the confusion matrix corresponding to the inference task at hand.

Confusion matrix formalisation

		Ground truth label	
		1	**0**
Prediction	**1**	True positive (TP)	False positive (FP)
	0	False negative (FN)	True negative (TN)

Positive predictive value TP / (TP + FP)	False discovery rate FP / (TP + FP)
False omission rate FN / (TN + FN)	Negative predictive value TN / (TN + FN)

True positive rate TP / (TP + FN)	False positive rate FP / (FP + TN)
False negative rate FN / (FN + TP)	True negative rate TN / (FP + TN)

Confusion matrix example

		Actual recidivism	
		Re-offended	**Did not re-offend**
Risk of recidivism	**High-risk**	True positive	False positive
	Low-risk	False negative	True negative

Example: positive predictive value: how many individuals *did re-offend among all individuals who* were *predicted as high-risk?*

Example: true positive rate: how many individuals were *predicted as high-risk* among all individuals who did *re-offend*?

We talk about a true positive inference (or label) when an individual who did recidivate is attributed a high-risk label, a a true negative when an individual who did not recidivate is attributed a low-risk label, a false positive when an individual who did not recidivate is incorrectly attributed a high-risk label, and a false negative when an individual who did recidivate is incorrectly attributed a low-risk label.

Usually, these inferences and their ground truth label (i.e. whether someone did recidivate or not) are summarised into a confusion matrix, as shown in Figure 4.

Machine learning models are usually evaluated using metrics that count various combinations of these inferences.

For instance, we usually talk about accuracy when counting the number of correct inferences, i.e. true positives and true negatives, among the total number of inferences (true positives and negatives, and false positives and negatives).

We talk about true positive rate when counting the number of true positive inferences (the individuals who were attributed a high-risk label and who did recidivate) over the total number of "positive" individuals, i.e. individuals who did recidivate.

While errors in machine learning inferences are unavoidable, points out that different stakeholders might value different types of errors, and hence might prefer different metrics.[244]

For instance, a decision-maker might want to know how many individuals will indeed recidivate (positive individuals, i.e. true positive) of those that have been labeled as high risk (true positive and false positive).

A defendant might want to know their probability to be incor rectly classified as high-risk, i.e. the ratio of false positives on the total number of individuals who did not positive and false positive).

A defendant might want to know their probability to be incorrectly classified as high-risk, i.e. the ratio of false positives on the total number of individuals who did not positive and false positive).

A.3 Warning: other machine learning "biases"

Machine learning researchers have also termed certain technical concepts as "bias" (e.g. overfitting, spurious correlations), without referring to social biases. We believe it is important for the reader to be aware of these other concepts to avoid any confusion when employing the word "bias" without a precise context.

Bias can refer to spurious correlations learned by a machine learning model. For instance, a model for classifying images of ambulances could have learned to identify solely the presence of a doctor and a flashing light in a picture to label an image as an ambulance, which this could be considered spurious in contexts where for instance other cars could also have flashing lights (e.g. firetruck), and where doctors could be present.

Bias can also relate to the concept of overfitting in machine learning.

Overfitting refers to when a model is trained on few data compared to the number of parameters it has, and consequently learns to infer the labels of the training data very accurately, whereas the training data might not fully be representative of the data seen at deployment time, and hence the model inferences are too specific and not necessarily accurate on the deployment data.

In statistics, a bias is the difference between an estimator's expected value and its true value, an estimator being a rule to calculate an estimate of a quantity based on sampled observations.

---

244 **Arvind Narayanan. 2018.** Translation tutorial: 21 fairness definitions and their politics. In Proc. Conf. Fairness Accountability Transp., New York, USA, Vol. 1170.

B. An introduction to the socio-technical notion of bias

Seeing the numerous ways in which the term bias is employed in scientific literature, as well as how it is conflated with other possibly similar terms like harm, diversity, discrimination, we judge it necessary to present an overview of the terms to the readers, for them to better understand where early computer science literature on debiasing stems from and grasp the scientific discourse; and the assumptions it makes.

The data mining community, which aims at extracting information from datasets, were first to use the term 'algorithmic discrimination' in computer science. Around 2008, these publications talk about discrimination, referring to civil rights laws "discrimination refers to unfair or unequal treatment of people based on membership to a category or a minority, without regard to individual merit."[245]

It then expanded to the field of machine learning around 2011, as machine learning techniques are progressively applied to the same datasets.

With this shift, there was also a shift in terminology, talking about fairness more than discrimination,[246] and taking inspiration from social choice theory and political philosophy (e.g. Dwork et al. discuss the notion of equality of opportunity).[247]

At this time, the term bias is used already in machine learning, without referring to social issues but to technical biases,[248] or to biases from annotators of datasets solely.[249]

Around 2015, bias starts to be used referring to biases in datasets and the harmful social issues they lead to. Then, the data management community also discusses issues of discrimination and responsible use of data, primarily referring to the "coverage" power of the data present in a database (e.g. all minorities are represented in sufficient quantity), and to "diversity" in the results of queries to database management systems.

In 2017, [Crawford [n.d.]] discuss the harms the use of biased systems cause, differentiating between allocation harms –the system would unfairly allocate resources to certain groups of population and not others–, and representation harms –the system stereotypically represent the identity of certain populations.

With this evolution of the terms, the critical discussion in industry has moved slowly from the collection of sensitive data in relation to privacy to the use of this data in machine learning systems. Fairness and discrimination issues indeed ask for regulation on the use of data, and often imply

the collection of additional data for mitigation, keeping out of the question the harms around the collection.

While debiasing is still under research, it is already operationalised within tools mostly stemming from industry in order to easily deploy "unbiased" systems.

[245] **Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008.** Discrimination-Aware Data Mining. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Las Vegas, Nevada, USA) (KDD '08). Association for Computing Machinery, New York, NY, USA, 560–568. https://doi.org/10.1145/1401890.1401959

[246] **Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011.** Fairness-aware learning through regularization approach. In 2011 IEEE 11th International Conference on Data Mining Workshops. IEEE, 643–650. https://www.kamishima.net/archive/2011-ws-icdm_padm.pdf

[247] **Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012.** Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference. 214–226. https://arxiv.org/abs/1104.3913

[248] **Antonio Torralba and Alexei A Efros. 2011**. Unbiased look at dataset bias. In CVPR 2011. IEEE, 1521–1528. https://ieeexplore.ieee.org/document/5995347

[249] **Trevor Cohn and Lucia Specia. 2013.** Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 32–42.; **Ece Kamar, Ashish Kapoor, and Eric Horvitz. 2015.** Identifying and accounting for task-dependent bias in crowdsourcing. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 3. https://aclanthology.org/P13-1004

References

Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. 2019. One-network adversarial fairness. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 2412–2420.

Philip Agre. 1997. Beyond the mirror world: Privacy and the representational practices of computing. Technology and privacy: The new landscape (1997), 29–61.

Yongsu Ahn and Yu-Ru Lin. 2019. Fairsight: Visual analytics for fairness in decision making. IEEE transactions on visualization and computer graphics 26, 1 (2019), 1086– 1095.

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-30.

Syed Mustafa Ali. 2018. AI and Epistemic Injustice: Whose Knowledge and Authority?

Aouragh, M., Gürses, S., Pritchard, H., & Snelting, F. (2020). The extractive infrastructures of contact tracing apps. Journal of Environmental Media, 1(2), 9-1.

Natã M Barbosa and Monchu Chen. 2019. Rehumanized crowdsourcing: a labeling framework addressing bias and ethics in machine learning. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–12.

Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. Calif. L. Rev. 104 (2016), 671.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623).

Berg, J. et al. (2018). Digital Labour Platforms and the Future of Work: Towards Decent Work in the Online World. At https://www.ilo.org/global/publications/books/WCMS_645337/lang--en/index.htm

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression.

Biega, Asia J., Krishna P. Gummadi, and Gerhard Weikum. "Equity of attention: Amortizing individual fairness in rankings." The 41st international acm sigir conference on research & development in information retrieval. 2018.

Binns, Reuben. "On the apparent conflict between individual and group fairness." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.

Birhane, Abeba, and Olivia Guest. "Towards decolonising computational sciences."

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach,and Kathleen Walker. [n.d.]. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report. Technical Report MSR-TR-2020-32, Microsoft, May 2020.

Matthias Boehm; Arun Kumar; Jun Yang, 2019. Data Management in Machine Learning Systems , Morgan & Claypool

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems 29 (2016), 4349–4357.

Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society 15, 5 (2012), 662–679.

Benedetta Brevini. 2020. Black boxes, not green: Mythologizing artificial intelligence and omitting the environment. Big Data & Society 7, 2 (2020), 2053951720935141.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency. 77–91.

Robin Burke. 2017. Multisided fairness for recommendation.

Burke, M., & Lobell, D. B. (2017). Satellite-based assessment of yield variation and its determinants in smallholder African systems. Proceedings of the National Academy of Sciences, 114(9), 2189-2194.

Ingrid Burrington, The Environmental Toll of a Netflix Binge, The Atlantic, 16. December 2015. https://www.theatlantic. com/technology/archive/2015/12/there-are-no-clean-clouds/420744

Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. In 2019 IEEE Conference on Visual Analytics Science and Technology (VAST). IEEE, 46–56.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356, 6334 (2017), 183–186.

Stevie Chancellor, Shion Guha, Jofish Kaye, Jen King, Niloufar Salehi, Sarita Schoenebeck, and Elizabeth Stowell. 2019. The Relationships between Data, Power, and Justice in CSCW Research. In Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing. 102–105.

Chasalow, Kyla, and Karen Levy. "Representativeness in Statistics, Politics, and Machine Learning." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.

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

Stop LAPD Spying Coalition. 2021. Stop LAPD Spying Coalition. https://stoplapdspying. org

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 32–42.

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

Hannah Couchman. 2019. Liberty's briefing on police use of live facial recognition technology. https://www.libertyhumanrights.org.uk/wp-content/uploads/2020/02/LIBERTYS-BRIEFING-ON-FACIAL-RECOGNITION-November-2019-CURRENT.pdf

Council of the EU. 2020. The charter of fundamental rights in the context of artificial intelligence and digital change.

Council of the EU. 2020. Preventing discrimination caused by the use of artificial intelligence.

Kate Crawford. [n.d.]. The Trouble With Bias. Keynote at NeurIps. 2017. Kate Crawford and Trevor Paglen. 2019. Excavating AI: The politics of images in machine learning training sets. Ex cavating AI (2019)

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

Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does object recognition work for everyone?. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 52–59.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 67–73.

Ravit Dotan and Smitha Milli. 2020. Value-laden disciplinary shifts in machine learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 294–294.

Dubal, V. B. (2017). The Drive to Precarity: A Political History of Work, Regulation, & Labor Advocacy in San Francisco's Taxi & Uber Economies. Berkeley Journal of Employment and Labor Law, 73-135.

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

Edelman, Benjamin, Michael Luca, and Dan Svirsky. "Racial discrimination in the sharing economy: Evidence from a field experiment." American Economic Journal: Applied Economics 9.2 (2017): 1-22.

EDRi. 2021. Civil society calls for AI red lines in the European Union's Artificial Intelligence proposal. https://edri.org/our-work/civil-society-call-for-ai-red-lines-in-the-european-unions-artificial-intelligence-proposal

Michael D Ekstrand, Robin Burke, and Fernando Diaz. 2019. Fairness and discrimination in retrieval and recommendation. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 1403–1404.

Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In Conference on Fairness, Accountability and Transparency. Pmachine learningR, 160–171.

Erikson and Lacerda. 2008. Charlatanry in forensic speech science, The International Journal of Speech, Language and the Law [IJSSL], vol 14.2, 169–193.

Virginia Eubanks. 2018. The digital poorhouse. Harper's Magazine (2018).

European Commission Data Protection Working Party (article 29). 2018. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation.

European Commission. 2020a. EU Anti-racism Action Plan 2020-2025.

European Commission. 2020b. White Paper on Artificial Intelligence: A European Approach to Excellence and Trust.

European Commission. 2021. 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.

European Digital Rights. 2020. Ban Biometric Mass Surveillance https://edri.org/wp-content/uploads/2020/05/Paper-Ban-Biometric-Mass-Surveillance.pdf

European Union Agency for fundamental rights. 2019. Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights.

Farkas, Lilla. "Analysis and comparative review of equality data collection practices in the European Union: Data collection in the field of ethnicity." Directorate-General for Justice and Consumers Directorate D–Equality Unit JUST D 1 (2017).

Sina Fazelpour and Zachary C. Lipton. 2020. Algorithmic Fairness from a Non-Ideal Perspective. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (New York, NY, USA) (AIES '20). Association for Computing Machinery, New York, NY, USA, 57–63. https://doi.org/10.1145/3375627.3375828

Fiebig, Tobias, Seda Gürses, Carlos H. Gañán, Erna Kotkamp, Fernando Kuipers, Martina Lindorfer, Menghua Prisse, and Taritha Sari. 2021. Heads in the Clouds: Measuring the Implications of Universities Migrating to Public Clouds.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 12.

Fredman, Sandra. "Intersectional discrimination in EU gender equality and non-discrimination law." European Commission, DG for Justice and Consumers, Directorate D–Equality, Unit JUST/DI, Brussels (2016). https://op.europa.eu/en/publication-detail/-/publication/d73a9221-b7c3-40f6-8414-8a48a2157a2f

Gandy Jr, O. H. (2021). The panoptic sort: A political economy of personal information. Oxford University Press.

Google. [n.d.]. Responsible AI practices. https://ai.google/responsibilities/responsible-ai-practices

Richard Gray. 2015. Flickr's autotag system mislabels concentration camps as 'jungle gyms'. https://www.dailymail.co.uk/sciencetech/article-3093074/Flickr-s-autotagturns-offensive-Image-recognition-software-mislabelsconcentration-campsjungle-gyms.htm

Nina Grgi□-Hla a, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive □ fairness in algorithmic decision making: Feature selection for procedurally fair learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32.

Sarah Griffiths, Why you internet use is not as clean as you think? BBB Smart Guide to Climate Change, 6. March 2020. https://www.bbc.com/future/article/20200305-why-your-internet-habits-are-not-as-clean-as-you-think

Seda Gurses and Joris Van Hoboken. [n.d.]. Privacy after the agile turn. ([n. d.]).

Thilo Hagendorff and Katharina Wezel. 2019. 15 challenges for AI: or what AI (currently) can't do. AI & SOCIETY (2019),1–11.

Isobel Asher Hamilton. 2019. Thousands of people across Europe are protesting and striking against Amazon on Black Friday. https://www.businessinsider.com/amazon-strikes-and-protests-sweep-across-europe-on-black-friday-2019-11?r=US&IR=T

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in al gorithmic fairness. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 501–512.

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

Harlan, S. L., Pellow, D. N., Roberts, J. T., Bell, S. E., Holt, W. G., & Nagel, J. (2015). Climate justice and inequality. Climate change and society: Sociological perspectives, 127-163.

Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 392–402.

Harvey, D. 2004. The 'new' imperialism: accumulation by dispossession. Socialist Register 40: 63-87.

Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, et al. 2018. Applied machine learning at facebook: A datacenter infrastructure per spective. In 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 620–629.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In European Conference on Computer Vision. Springer, 793–811.

Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse.Information, Communication & Society 22, 7 (2019), 900–915.

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

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. (2020).

Elle Hunt. 2016. Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

IBM. [n.d.]. IBM's multidisciplinary, multidimensional approach to AI ethics. https://www.ibm.com/artificial-intelligence/ethics

Privacy International. 2020. The SyRI case: a landmark ruling for benefits claimants around the world. https://privacyinternational.org/news-analysis/3363/syri-case-landmark-ruling-benefits-claimants-around-world

Irani, L. (2015a). Justice for 'Data Janitors'. At www.publicbooks.org/justice-for-data-janitors

Michael Jackson. 1995. The world and the machine. In 1995 17th International Conference on Software Engineering. IEEE, 283–283.

Abigail Z Jacobs and Hanna Wallach. 2019. Measurement and fairness.

Jamil, R., & Noiseux, Y. (2018). Shake that moneymaker: insights from Montreal's Uber drivers. Revue Interventions Économiques. Papers in Political Economy, (60)

Ece Kamar, Ashish Kapoor, and Eric Horvitz. 2015. Identifying and accounting for task-dependent bias in crowdsourcing.In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 3.

Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination.Knowledge and Information Systems 33, 1 (2012), 1–33.

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

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In 2011 IEEE 11th International Conference on Data Mining Workshops. IEEE, 643–650.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 3819–3828.

Nicolas Kayser-Bril. 2020. Google apologizes after its Vision AI produced racist results. https://algorithmwatch.org/en/google-vision-racism

Kearns, Michael, et al. "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness." International Conference on Machine Learning. Pmachine learningR, 2018.

Os Keyes, Jevan Hutson, and Meredith Durbin. "A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry." Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. 2019.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. 43–53.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Nick Kolakowski, Google's Duplex Evolving: Can it meet your needs?, Dice, 8. May 2019. https://insights.dice.com/2019/05/08/googles-duplex-evolving-can-it-meet-your-needs

Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020. POTs: protective optimization technologies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 177–188.

Arun Kumar, Matthias Boehm, and Jun Yang. 2017. Data management in machine learning: Challenges, techniques, and systems. In Proceedings of the 2017 ACM International Conference on Management of Data. 1717–1722.

Kusner, Matt, et al. "Counterfactual fairness." Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2019. Delayed impact of fair machine learning. In Proceedings of the 28th International Joint Conference on Artificial Intelligence. AAAI Press, 6196–6200.

Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. 2019. Bias mitigation post-processing for individual and group fairness. In Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2847–2851.

Alan Lundgard. 2020. Measuring Justice in Machine Learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) (FAT* '20).

Association for Computing Machinery, New York, NY, USA, 680. https://doi.org/10.1145/3351095.3372838

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. On the Applicability of machine learning Fairness Notions.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning.

Jacob Metcalf and Kate Crawford. 2016. Where are human subjects in big data research? The emerging ethics divide. Big Data & Society 3, 1 (2016), 2053951716650211.

Microsoft. [n.d.]. Microsoft AI principles. https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6

Milli, Smitha, et al. "The social cost of strategic classification." Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application 8 (2021).

Petra Molnar 2020. Technological Testing Grounds: Migration Management and Reflections from the Ground Up. (published by EDRi) https://edri.org/our-work/technological-testing-grounds-border-tech-is-experimenting-with-peoples-lives

Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. 2019. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 1–36.

Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In Proc. Conf. Fairness Accountability Transp., New York, USA, Vol. 1170.

Northpointe. 2011. Compas Risk Assessment CORE. https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.htm

Rebekah Overdorf, Bogdan Kulynych, Ero Balsa, Carmela Troncoso, and Seda Gürses. 2018. Questioning the assumptions behind fairness solutions.

Praveen Paritosh, Panos Ipeirotis, Matt Cooper, and Siddharth Suri. 2011. The computer is the new sewing machine: benefits and perils of crowdsourcing.
In Proceedings of the 20th international conference companion on World wide web. 325–326.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2799–2804.

Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. 2020. FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. In Proceedings of The Web Conference 2020. 1194–1204.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-Aware Data Mining. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Las Vegas, Nevada, USA) (KDD '08).

Association for Computing Machinery, New York, NY, USA, 560–568. https://doi.org/10.1145/1401890.1401959

Seeta Peña Gangadharan and J drzej Niklas. 2019. Decentering technology in discourse on discrimination. Information, Communication & Society 22, 7 (2019), 882–899.

Polyzotis, Neoklis, et al. 2017. Data management challenges in production machine learning. Proceedings of the 2017 ACM International Conference on Management of Data.

Julia Powles. The Seductive Diversion of 'Solving' Bias in Artificial Intelligence. https://onezero.medium.com/theseductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision?

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2019. Assessing gender bias in machine translation: a case study with google translate. Neural Computing and Applications (2019), 1–19.

Propublica. [n.d.]. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Prug, T.; Bilić, P., Work Now, Profit Later: AI Between Capital, Labour and Regulation Augmented Exploitation: Artificial Intelligence, Automation, and Work

Moore Phoebe V. ; Woodcok, Jamie (ur.). London, UK: Pluto Press, 2021. str. 30-40

Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 469–481.

Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 429–435.

Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 145–151.

Rieke, A., Yu, H., Robinson, D., & Van Hoboken, J. (2016). Data brokers in an open society.

Sadie Robinson. 2020. Furious students protest over A-Levels scandal. https://socialistworker.co.uk/art/50488/Furious+students+protest+over+A+Levels+scandal

Joel Ross, Lilly Irani, M Six Silberman, Andrew Zaldivar, and Bill Tomachine learninginson. 2010. Who are the crowdworkers? Shifting demographics in Mechanical Turk. In CHI'10 extended abstracts on Human factors in computing systems. 2863–2872.

Bonnie Ruberg and Spencer Ruelos. 2020. Data for queer lives: How LGBTQ gender and sexuality identities challenge norms of demographics. Big Data & Society 7, 1 (2020), 2053951720933286.

Debjani Saha, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, and Michael Tschantz. 2020. Measuring non-expert comprehension of machine learning fairness metrics. In International Conference on Machine Learning, 8377–8387.

Galhotra, Sainyam, Yuriy Brun, and Alexandra Meliou. "Fairness testing: testing software for discrimination." Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering.

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

Javier Sánchez-Monedero and Lina Dencik. 2020. The politics of deceptive borders: 'biomarkers of deceit' and the case of iBorderCtrl. Information, Communication & Society (2020), 1–18.

Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. 2020. What does it mean to 'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 458–468.

Sapieżyński, Piotr, et al. "Algorithms that" Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences."

Sapieżyński, Piotr, et al. "Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists." Companion Proceedings of The 2019 World Wide Web Conference.

Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2020. How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. Artificial Intelligence 283 (2020), 103238.

Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. 2019. FairPrep: Promoting Data to a First-Class Citizen in Studies on Fairness-Enhancing Interventions.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency. 59–68.

Eunjin Seong and Seungjun Kim. 2020. Designing a Crowdsourcing System for the Elderly: A Gamified Approach to Speech Collection. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 1–9.

Singh, Ashudeep, and Thorsten Joachims. "Fairness of exposure in rankings." Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018.

Singh, Harvineet, et al. "Fairness violations and mitigation under covariate shift." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

Olivia Solon, Big Tech call center workers face pressure to accept home surveillance, NBC News, https://www.nbcnews.com/tech/tech-news/big-tech-call-center-workers-face-pressure-accept-homesurveillance-n1276227

Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2459–2468.

Stark, Luke, and Jesse Hoey. "The ethics of emotion in artificial intelligence systems." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.

Julia Stoyanovich, Bill Howe, and HV Jagadish. 2020. Responsible data management. Proceedings of the VLDB Endowment 13, 12 (2020), 3474–3488.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. Association for Computational Linguistics (ACL 2019) (2019).

Sühr, Tom, et al. "Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019.

Syed Mustafa Ali. A brief introduction to decolonial computing. XRDS: Crossroads, The ACM Magazine for Students, 22(4):16–21, 2016.

Taylor, K. Y. (2019). Race for profit: How banks and the real estate industry undermined black homeownership. UNC Press Books.

Taylor, L. (2021). Public actors without public values: legitimacy, domination and the regulation of the technology sector. Philosophy & technology, 1-26.

Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In CVPR 2011. IEEE, 1521–1528.

Zeynep Tufekci. 2015. Algorithms in our midst: Information, power and choice when software is everywhere. In Proceedings of the 18th acm conference on computer supported cooperative work & social computing. 1918–1918.

UN. 2020. New information technologies, racial equality, and non-discrimination: Call for input. https://www.ohchr.org/EN/Issues/Racism/SRRacism/Pages/Info-Technologies-And-Racial-Equality.aspx

Ustun, Berk, Alexander Spangher, and Yang Liu. "Actionable recourse in linear classification." Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019.

Van Doorn, N., Ferrari, F., & Graham, M. (2020). Migration and migrant labour in the gig economy: an intervention. Available at SSRN 3622589.

Van Hoboken, Joris. "From collection to use in privacy regulation? A forward-looking comparison of European and us frameworks for personal data processing." Exploring the Boundaries of Big Data 231 (2016).

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

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

Jennifer Wortman Vaughan, 2017. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research, 18(193):1□46, 2018.

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

Waldman, A. E. 2019. Power, process, and automated decision-making. Fordham L. Rev., 88, 613. Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In Proceedings of the IEEE Interntional Conference on Computer Vision. 5310–5319.

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

Patrick Williams and Eric Kind. 2019. Data-driven Policing: The hardwiring of discriminatory policing practices across Europe. (published by the European Network Against Racism) https://www.enar-eu.org/IMG/pdf/data-drivenprofiling- web-final.pdf

Ben Williamson and Anna Hogan. 2020. Commercialisation and Privatisation in/of Education in the Context of COVID-19. Tech. rep. Education International.

Yaseen Aslam and Jamie Woodcock.2020. A History of Uber Organizing in the UK. http://oro.open.ac.uk/71933

Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In Proceedings of the 2018 CHI conference on human factors in computing systems. 1–14.

Wenying Wu, Pavlos Protopapas, Zheng Yang, and Panagiotis Michalatos. 2020. Gender Classification and Bias Mitigation in Facial Images. In 12th ACM Conference on Web Science. 106–114.

Huichuan Xia, Yang Wang, Yun Huang, and Anuj Shah. 2017. " Our Privacy Needs to be Protected at All Costs" Crowd Workers' Privacy Experiences on Amazon Mechanical Turk. Proceedings of the ACM on Human-Computer Interaction 1, CSCW (2017), 1–22.

Alice Xiang and Inioluwa Deborah Raji. 2019. On the Legal Compatibility of Fairness Definitions.

Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. 2019. Balanced Ranking with Diversity Constraints. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 547–558.

Karen Yeung, PE Vermaas, and I van de Poel. 2015. Design for the Value of Regulation. Handbook of Ethics, Values, and Technological Design (2015), 447–472.

Ming Yin, Siddharth Suri, and Mary L Gray. 2018. Running Out of Time: The Impact and Value of Flexibility in On-Demand Crowdwork. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 1–11.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In Artificial Intelligence and Statistics. Pmachine learningR, 962–970.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In International Conference on Machine Learning. 325–333.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang Chang. 2018. Learning Gender-Neutral Word Embeddings. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.

Kathryn Zyskowski, Meredith Ringel Morris, Jeffrey P Bigham, Mary L Gray, and Shaun K Kane. 2015. Accessible crowdwork? Understanding the value in and challenge of microtask employment for people with disabilities. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing.

Press enquiries
press@edri.org

Brussels office
brussels@edri.org

Phone number
+32 2 274 25 70

Visit us
Rue Belliard 12
1040 Brussels
Belgium

Follow us
Twitter
Facebook
LinkedIn
Youtube

EDRi

EUROPEAN DIGITAL RIGHTS

European Digital Rights (EDRi) is the biggest European network defending rights and freedoms online. We promote, protect and uphold human rights and the rule of law in the digital environment, including the right to privacy, data protection, freedom of expression and information.

www.edri.org