Fairness in agreement with European values
An interdisciplinary perspective on ai regulation

Colmenarejo, Alejandra Bringas; Nannini, Luca; Rieger, Alisa; Scott, Kristen M.; Zhao, Xuan; Patro, Gourab K.; Kasneci, Gjergji; Kinder-Kurlanda, Katharina

DOI
10.1145/3514094.3534158

Publication date
2022

Document Version
Final published version

Published in
AIES 2022 - Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society

Fairness in Agreement With European Values: An Interdisciplinary Perspective on AI Regulation

Alejandra Bringas Colmenarejo
University of Southampton
United Kingdom

Luca Nannini
Minsait - Indra Sistemas
CiTIUS, Universidade de Santiago de Compostela
Spain

Alisa Rieger
Delft University of Technology
Netherlands

Kristen M. Scott
KU Leuven
Belgium

Xuan Zhao
SCHUFA Holding AG
University of Tuebingen
Germany

Gourab K. Patro
IIT Kharagpur, India
L3S Research Center, Germany

Gjergji Kasneci
SCHUFA Holding AG
University of Tuebingen
Germany

Katharina Kinder-Kurlanda
Digital Age Research Center
University of Klagenfurt
Austria

ABSTRACT

With increasing digitalization, Artificial Intelligence (AI) is becoming ubiquitous. AI-based systems to identify, optimize, automate, and scale solutions to complex economic and societal problems are being proposed and implemented. This has motivated regulation efforts, including the Proposal of an EU AI Act. This interdisciplinary position paper considers various concerns surrounding fairness and discrimination in AI, and discusses how AI regulations address them, focusing on (but not limited to) the Proposal. We first look at AI and fairness through the lenses of law, (AI) industry, sociotechnology, and (moral) philosophy, and present various perspectives. Then, we map these perspectives along three axes of interests: *(i) Standardization vs. Localization*, *(ii) Utilitarianism vs. Egalitarianism*, and *(iii) Consequential vs. Deontological ethics* which leads us to identify a pattern of common arguments and tensions between these axes. Positioning the discussion within the axes of interest and with a focus on reconciling the key tensions, we identify and propose the roles AI Regulation should take to make the endeavor of the AI Act a success in terms of AI fairness concerns.

CCS CONCEPTS

• **Social and professional topics → Governmental regulations**; • **Applied computing → Law**.

KEYWORDS

AI Regulation, EU AI Proposal, Deontological Ethics, Consequential Ethics, Utilitarian Welfare, Egalitarian Welfare, Localization, Standardization

ACM Reference Format:
Alejandra Bringas Colmenarejo, Luca Nannini, Alisa Rieger, Kristen M. Scott, Xuan Zhao, Gourab K. Patro, Gjergji Kasneci, and Katharina Kinder-Kurlanda. 2022. Fairness in Agreement With European Values: An Interdisciplinary Perspective on AI Regulation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3514094.3534158

1 INTRODUCTION

AI applications have grown at an unprecedented rate in recent years and have become ubiquitous in our society. While often deployed with the intention to increase efficiency and fairness of decision-making, AI has also sparked many debates on (un)fairness [101]. These debates surround, amongst others, unfair treatment of individuals and groups due to the reproduction of systemic, institutional, and societal biases in AI decisions [14]; the opacity of AI decisions [4]; diverse jeopardies to democracy and societal well-being [74]; risks to consumer privacy [63]; and market inequalities that are observed in the aggregation of unprecedented levels of power of big companies that develop AI systems *(Big Tech)* while small and new companies are struggling to enter the market [100]. In many fields of AI application, such as policing, justice, and recruitment, bias and unfairness as described above should not only be mitigated to increase fairness but in fact, to avert violating protected human rights.

The above mentioned undesired effects and consequences of AI application and development propelled the European Union for new regulations, ex-ante reviews, and ex-post monitoring on AI systems. The European Union intends to assert the AI Regulation through the protection of human dignity and fundamental rights with the

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08…$15.00
https://doi.org/10.1145/3514094.3534158

Proposal of the *Artificial Intelligence Act* [85], convinced that human beings should remain at the center of technological development. However, to make this endeavor of the AI Act a success, to some extent divergent interdisciplinary views and perspectives on bias, fairness, and regulation, have to be taken into consideration.

We elaborate on *legal*, *industrial*, *sociotechnical*, and *philosophical* perspectives in light of identified axes of tension in the debate on AI fairness and regulation: *Standardization vs. Localization*, *Utilitarianism vs. Egalitarianism*, and *Consequential vs. Deontological.* Further, we discuss discrepancies between how these perspectives are addressed in the current Proposal of the Artificial Intelligence Act and make recommendations how they could be addressed for better reconciliation with all three perspectives and the legal requirements. In sum, we make the following contributions to the ongoing discourse on AI fairness and regulation: i. **Interdisciplinary perspectives:** Comprehensive interdisciplinary (technical, legal, industrial, sociotechnical, philosophical) discussion of bias, fairness, and regulation (sections 2 to 6), ii. **Mapping tensions of debate:** mapping the different perspectives on fairness in AI applications and regulation on to three axes that reveal tensions in the debate: *Standardization vs. Localization*, *Utilitarianism vs. Egalitarianism*, and *Consequential vs. Deontological ethics* (section 7), iii. **Path forward:** Recommendations towards consensus for a successful AI Act that reconciles divergent perspectives (section 8).

2 TECHNICAL FRAMEWORKS FOR BIAS AND FAIRNESS IN AI

In this section we present examples of fairness controversies for selected AI application domains with high-stake consequences. Subsequently, we discuss several AI fairness notions and present research on guidance to choose between these notions and between measures to mitigate bias in AI systems.

2.1 Examples of Bias and Unfairness in AI Applications

Automated decision-making systems were suggested to be capable of increased fairness due to avoidance of human bias interference [52]. However, many cases have come to light in which automatic decision-making was found to raise critical issues regarding fairness, and reproduces systemic, institutional, and societal biases. Such biases can result in discrimination, unfairness, and issues of privacy, thus, violating protected human rights (see section 3). This is especially harmful when automated decision making has high-stake implications for individuals and society. In the following, we present salient examples.

In **Policing and Justice**, AI systems are applied across Europe to inform and assist day-to-day police work by profiling people, attempting to predict likely future behavior or locations of future crimes, and assessing the alleged risk of criminal involvement of individuals (e.g., *Top 600 criminals list* and *CAS* (Netherlands), *Delia* (Italy), *SKALA* (Germany). Outcomes of these predictions and assessments are used to justify surveillance, searches, or questioning of alleged *high risk* individuals. However they have been suspected to reinforce existing patterns of offending and enforcement [1, 99]. In the judicial arena, automated decision-making is currently being applied in various courts around the world to support certain tasks, such as risk assessment of recidivism, as well as decisions concerning bail amounts, probation periods, and sentencing [94, 116]. Across Europe, such systems are not yet used widely, however, they have been introduced or tested in some countries, e.g., in Spain *(RisCanvi)* or the UK *(HART)*. Završnik [116] highlights potentially violated rights due to opaque, automated decision-making in the justice system, e.g., the right to a fair trial, the principle of non-discrimination and equality, and the right for explanation.

AI systems are further being applied in the domain of **Education and Employment**, to support candidate selection for higher education admissions and recruitment, e.g., with CV screening, targeted job advertisement, candidate sourcing, and video screening [2]. The risk of bias has been demonstrated at each of these stages in the recruitment process [11, 56].

In **Finance and Banking**, AI algorithms constitute the basis of numerous different applications, such as market forecasting for trading, or risk management for credit scoring, loan allocations, and mortgage rates [15]. Various cases have come to light in which decisions of such applications were found to be unfair and biased towards minority borrowers, i.e., with higher mortgage and loan rejection rates for Hispanic and Black borrowers in the US [8, 29], or lower credit limits for women than for men with equal credit relevant characteristics [36, 106].

For **Online Platforms**, AI based recommender systems are applied to support users to navigate the web by filtering information and suggest items (videos, social media content, products, music,..) predicted to be relevant for the user. Recommender systems were found to amplify different kinds of bias, such as *representation* bias with an over-representation of male, white, and young users [96], and *exposure* bias where the top 20% of businesses get 80% of the exposure [87], and marketplaces preferentially recommend their own products [21]. This amplifies substantial power imbalances between market-dominating platform incumbents *(Big Tech)* and smaller platforms who do not have access to equal vast amounts of high-quality consumer data that is vital to enter the market [100]. The resulting immense power concentration in the private hands of very few companies that develop most AI applications and prioritize profit over benevolence for society poses an additional threat to democracy and society [28, 105]. Further, recommender systems and search result rankings that often optimize to capture attention, determine a large extent of the information to which people are exposed. This can result in distorted exposure to information and viewpoints, as well as exposure to dis- and misinformation, raising issues of fairness and posing a threat to democracies that are reliant on well-informed citizens who can engage in healthy political and social discourse [42, 74]. AI systems could threaten democracy and society further by undermining the process of elections through targeted advertisements. Such *microtargeting* provides tools for interference by malicious political actors [23, 73].

2.2 Mitigating Bias and Ensuring Fairness

Most fairness definitions consider either group or individual fairness. *Group fairness* is focused on requiring that people who belong to protected groups receive on average the same treatment/outcome as the overall population, expressed as the equality of a selected statistical measure across groups [111], such as *statistical parity,* *demographic parity, equal opportunity* and *equality of odds. Individual fairness* focuses on ensuring that any two individuals who are similar except for the protected features receive equal or similar treatment/outcomes [26]. While ideally, multiple fairness notions would be met to reach a *complete* fairness status, this is impossible due to mathematical incompatibilities between them [68]. Criteria to systematize the procedure of selecting between fairness notions when making a specific decision have been proposed: Amongst others, the existence of a ground-truth, base-rates between subgroups, the cost of misclassification, or the existence of government regulations to meet may be considered [62].

Formalization of fairness definitions in a specific context is nuanced and it is important that AI practitioners receive some guidance when designing a fair AI system. Some recent research proposes the *Fairness Compass*, a schema in form of a decision tree which simplifies the selection process by settling for the desired ethical principles in a formalised way [98]. A *standardized roadmap* could potentially make the identification of an appropriate fairness definition a more straightforward procedure, and help document the decision process toward fairness. Audit, monitoring and explanation might then be more accessible and less expensive. Nevertheless, there should also be space for stakeholders with deeper understanding of the specific context to contribute refinement and interpretations of any such roadmap.

The fairness notions mentioned above deal with the outcome of automated decision-making. counterfactual fairness [55] and causal fairness [112], however, have a procedural implication which might be more suitable for the cases where a counterfactual or causal connection needs to be established between features. Most of the existing fairness notions are formalized in a static scenario. If we want to better understand how bias is encoded in historical data or evaluate the consequences of certain fairness intervention, dynamic fairness notions [20] might offer a better solution.

Technical methods to mitigate bias in algorithms fall under three categories: (1) *Pre-processing*. Pre-processing techniques try to transform/re-balance the data so that the underlying discrimination is mitigated; (2) *In-processing*. The construction of objective function usually has Utilitarian motivation behind, e.g. trying to maximize the utility of whole population. In-processing methods for bias mitigation can be used either by incorporating changes into the objective function or imposing a fairness constraint; (3) *Post-processing*. Post-processing methods reassign the labels initially predicted by the black-box model to a fairer state. [66].

The existing technical solutions toward fairness focus on more consequential approaches: the outcome/decision is evaluated by a specific fairness notion and then measures are taken to correct the unfair outcome/decision. Concerns have been voiced that fairness cannot be simply achieved through mathematical formulation approaches as the *formalism trap* [103] and the seeming success of these technical solutions in the end will hinder pursuits of actual fairness with the cooperation of social practices [39].

3 A LEGAL PERSPECTIVE ON BIAS AND FAIRNESS IN AI

To follow one central goal of the EU—the promotion of peace and well-being for its members—EU law aims at ensuring that EU member-states and individuals are treated and treat each other equally and fairly. The blindfolded *Justicia* further emphasizes the importance of laws that promote fairness, but also fairness within the enforcement of all laws. Decision-making based on machine-learning could be a promising support for that, to mitigate the unconscious or deliberate biases that we as humans have. However, being trained on (biased) data from previous decisions, the promise of unbiased assessments could not be fulfilled so far [5, 46].

In this section, we will take a structured look at the legal perspective on bias and fairness in AI. We will start with an overview of EU legislative framework on non-discrimination and the approach to fairness followed by the EU Data Protection Law. Then we will conclude by addressing the technical requirements to deal with bias that would be introduced with the AI Regulation Proposal.

3.1 Non-Discrimination Law

The general principle of non-discrimination in EU law protects people from discrimination and unfair treatment. European anti-discrimination law is designed to prevent discrimination against particular groups of people that share one or more characteristics—called protected attributes—and from which the group acquires the category of a protected group. Concretely, protected attributes under the Charter of Fundamental Rights of the European Union include sex, race or ethnic origin, colour, ethnic or social origin, genetic features, religion or other belief, disability, age, sexual orientation, political or any other opinion, language, membership to a national minority, property, social origin, and birth (Art. 21.(1)) [82]. Additionally, the Charter prohibits discrimination on the grounds of nationality, compels the European Union to ensure the equality of everyone under the European law, demands the respect of cultural, religious, and linguistic diversity, and seeks equality of men and women in all areas. Several other European anti-discrimination directives have further covered the legal protection offered to these protected attributes. Specifically, under the European Legislation men and women must receive equal treatment in the labour market and regarding the access and supply of good as services[79, 83]. Likewise, equal treatment must be guaranteed between persons irrespective of their racial or ethnic origin [78], as well as equity shall be respected in employment and occupation in regards to the grounds of disability, religion or belief, age and sexual orientation [77]. Member States expanded the protection towards discrimination through specific national laws and provisions.

Furthermore, the European legislation presents two tools to address discrimination, *direct* and *indirect* discrimination. Direct discrimination is defined as a *situation in which one person is treated less favourable on the grounds of a prohibited criterion than another is, has been or would be treated in a comparable situation* [78]. Thus, it is straightforwardly related to the possession of a protected attribute that distinguishes the person from other individuals, regardless of the intention behind the disparate treatment or the mere existence of less favourable treatment. In the context of data-driven systems, direct discrimination will cover those cases where the model is not neutral towards a protected attribute and offers a less favourable output to individuals on the basis of protected groups, whether they truly fit into that group or are associated with the protected attribute. Since consciously inputting discrimination into the model will affect its accuracy, these cases are not of great concern [113].

By contrast, indirect discrimination will more likely capture many situations of algorithmic discrimination because it affects situations *where an apparently neutral provision, criterion or practice would put members of a protected category at a particular disadvantage compared with other persons unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary* [78]. Nevertheless, the prohibition of indirect discrimination does not encompass a set of clear and easily applicable rules, it can rather be considered closer to a standard than to a rule [118]. *The concept of indirect discrimination results in rather open-ended standards, which are often difficult to apply in practice. It needs to be proven that a seemingly neutral rule, practice or decision disproportionately affects a protected group* [118]. Due to this, indirect discrimination concerns neutral models, which in principle are blinded to sensitive attributes or do not operate on the basis of those protective attributes. Thus, direct discrimination focuses on individual cases of discrimination, while indirect discrimination deals with rules and patterns of discrimination and can reveal underlying social inequalities.

3.2 Data Protection Law

The European Union General Data Protection Regulation (GDPR) [84] refers to automated individual decision-making and seeks, amongst other objectives, to prevent algorithmic discrimination. Generally, the GDPR states the objective to protect all the fundamental rights recognised under EU law, which the processing of personal data may challenge. According to the GDPR, the core principles that shall lead the processing of personal data are lawfulness, fairness, and transparency. Concretely, the principle of fairness entails the processing of personal information that is not in any way *unduly detrimental, unexpected, or misleading to the individuals concerned* ([48]). Indeed, the principle of fairness seeks to protect the individual's fundamental rights and freedoms, and so, their non-infringement by such processing. Likewise, the *principle of data accuracy* requires the control of the quality of data for its processing, although it does not address the possible wrongful or disproportionate selection of data and therefore the effect and consequences resulted from such selection [76].To ensure fair processing, the GDPR requests the use of *appropriate mathematical and statistical procedures for profiling that take into account the risks involved for the interest and rights of data subjects and prevent discriminatory effects on natural persons* (Recital 71 [84]). Furthermore, the GDPR highlights the potential *risks to the rights and freedom of natural persons, which could lead to physical, material or non-material damage, in particular when processing results in discrimination* (Recital 75 [84]). Despite these provisions, ensuring fairness is still quite a subjective matter as it requires that the data processing shall not exceed reasonable expectations nor provoke unjustified adverse effects on the individuals. However, what can be considered reasonable expectations and justifiable effects is an open question, leaving the notion of *fair processing* undefined.

However, the European anti-discrimination law evidently embedded notions of substantive discrimination and therefore, unjustified algorithmic discrimination, as referred to in Article 5 and Recital 71, implies unfair processing [38]. From the legal perspective, discrimination collides with equality, infringing the principle of fairness; whereas from a technical perspective, algorithmic discrimination straightforwardly entails unfair processing (see section 2).

3.3 EU Artificial Intelligence Regulation Proposal

With the EU Artificial Intelligence Act the European Union aims at laying down harmonized rules on artificial intelligence with four specific objectives [85]: *1) ensure that AI systems placed on the Union market are safe and respect existing law on fundamental rights and Union values; 2) ensure legal certainty to facilitate investment and innovation in AI; 3) enhance governance and effective enforcement of existing law and safety requirements applicable to AI systems; 4) facilitate the development of a single market for lawful, safe and trustworthy AI applications preventing market fragmentation.*

In essence, the Proposal seeks to balance legal certainty and the development of AI systems while ensuring an approach that respects European values, principles and laws. The specific purpose of the Proposal is to establish a classification for trustworthy AI systems based on a risk-based approach, to introduce new legal obligations and requirements on public authorities and businesses for the development and application of AI systems, to prohibit harmful AI-enabled practices, and to set new monitoring and enforcement regimes. Essentially, the Proposal will set a legal framework applicable for developers and end-users of AI systems which *specific characteristics—opacity, complexity, dependency on data, autonomous behaviours—can adversely affect a number of fundamental rights enshrined in the EU Charter of Fundamental Rights* [85].

The Proposal delimits a set of prohibited AI practices considered harmful because they contravene EU values and violate fundamental rights. Second, the Proposal outlines specific obligations to avoid the appearance of bias in two types of high-risk AI systems; (1) those which are intended to be used as a safety component of a product or is itself a product, and this product is subject to an existing third-party conformity assessment, and (2) those which are involved in decision-making processes in the following areas; (i) biometric identification and categorization of natural persons, (ii) management and operation of critical infrastructure, (iii) education and vocational training, (iv) employment and workers management as well as access to self-employment, (v) law enforcement, (vi) migration, asylum, and border control management, and (vii) administration of justice and democratic processes (see section 2.1).

According to the Proposal, AI systems can only be placed into the EU market if they comply with the certain minimum requirements specified in the legislation, requirements that become stricter as the risk associated with the system increases (i.e., minimal risk, low risk, high risk, and unacceptable risk). Consequently, providers will need to carry out ex-ante conformity assessments and implement quality and risk management systems and post-market monitoring to ensure compliance with the new regulation and minimise the risk for users and affected persons. However, the Proposal pays little attention to identifying the causes and proposing recommendations to tackle the potential discriminatory harms of AI systems. Specifically, the Proposal mainly focuses on biases in data sets, forgetting other types such as those that may arise from the choice of algorithms, and the optimization or evaluation of metrics. Additionally, the Proposal may pose unreasonable trust in human operators—i.e., human in the loop—to identify and recognise cases of bias and discrimination in AI systems.

The Proposal does not provide detailed guidance on dealing with unavoidable trade-offs for the different stakeholders when debiasing and monitoring bias in the data set. Nevertheless, some insights can be found in the Proposal regarding the expected requirements to debias high-risk AI systems. Firstly, there will be an obligation to establish appropriate data governance and management practices concerning the training, validation, and testing of data sets, in particular, to examine possible biases, ensure the relevance, representativeness, absence of errors and completeness of the data sets, and their consideration with the characteristics or elements that are particular to the specific geographical, behavioural or functional setting within which the high-risk AI system is intended to be used [85]. Secondly, a novel exception to the Data Protection Regulation will allow *to the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems* [85] the processing of special categories of data. Finally, the Proposal asks for developing methods that will ensure the detection of biased outputs and the consequent introduction of appropriate mitigation measures as it recognises the potential of AI systems to develop biased outputs due to outputs used as an input for future operations, i.e., *feedback loops*.

Interestingly, the Proposal also details the role of standards and specifications in the AI landscape [85]. On the one hand, the Proposal addresses the use of *harmonised standards* to presume the conformity of AI systems with the regulation's requirements. On the other hand, the Proposal entitles the Commission with the duty to adopt common specifications and technical solutions when the harmonised standards are insufficient or there is a need to address specific or fundamental rights concerns. In other words, *conformance with technical standards and common specifications should give providers of high-risk AI a level of confidence that they are compliant with the mandatory requirements of the proposed EU AI Regulation as well as significantly cutting the cost of compliance for business* [65]. Whereas neither the standards nor the specifications will be compulsory for providers of high-risk AI systems, their non-adoption shall entail a justification as to which and why other technical solutions were adopted.

4 AN INDUSTRY PERSPECTIVE ON BIAS AND FAIRNESS IN AI

Substantial research on ML fairness, even for industry applications, has originated out of academic contexts. Academic research has first proposed most fairness principles and quantitative methods to mitigate biases and unbalanced data with general application domains [6, 59, 66]. Toolkits appeared ready to be integrated for the industry, even if often developed following non-contextual design rationales based upon the issues of algorithmic methods [43]. Until recently, the technical nature of academic contributions have often not addressed the practical issues that industry practitioners face when adopting and engaging with fairness tools. Practitioners have pointed out the lack of ethical tools' usability in real-world applications due to a series of critical factors preventing the straightforward adoption of fairness principles and methods [69]. Following Morley et al. [71], such non-effectiveness in real-world cases stems from how fairness compliance is operationalized inside companies. If not developed with the sociotechnical features and constraints of AI product deployment in mind, these methods could easily lead to failures [43] including for example fairness definitions misinterpretation [54], obfuscation of practitioners' accountability [81], and gaming fairness measures as a method of ethics-washing [71]. To avoid shortcomings, researchers are now focusing on how to operationalize fairness frameworks based on the needs of industry practitioners. Veale et al. [110] conducted interviews with decision makers in high-stakes public-sector contexts. Practitioners were found to be lacking incentives and practices for algorithmic accountability due to resource constraints and dependency on prior infrastructure. Holstein et al. [44] enlarged the pool of industry practitioners with a systematic investigation of ML product development. Amid the area of intervention were identified issues of data quality provenance and reporting, as well as the need for domain-specific educational resources and compliance protocols, intended specifically as internal auditing processes and tools for fairness-focused debugging. Rakova et al. [92] reported that practitioners often felt a hostile organizational environment where they were hindered or uncompensated when trying to implement fairness practices independently. Disincentive stems from the lack of educational programs, rewards, accountability allocation, and communicative protocols over fairness issues, especially when different parts of an AI development are distributed across different teams. This resulted in practitioners often feeling disoriented, unprepared, or even overwhelmed by fairness tools and checklists [19, 44]. It was also observed that practitioners recommend establishing internal and external investigation committees to create an inclusive and preventive environment and to provide resources such as protocols or educational teams [61, 92]. Other research examples, once informed on practitioners' needs, focused on designing different AI fairness solutions: checklists to be aligned with teams' workflows and organizational ad-hoc processes, fairness frameworks or internal algorithmic auditing protocols designed for industrial applications [61, 91]. Recently, Richardson and Gilbert [97] proposed a complete industry framework of stakeholders and fairness recommendations while specifying operationalization pitfalls. Ibáñez and Olmeda [47] distinguished two main perspectives on operationalizing fairness practices in organizations: a bottom-up, reactive approach, where prior organizational processes restrain best practices, or top-down, where a proactive approach is set in place according to the translation of principles and methods as actionable, iterative steps designed with stakeholders' needs and concerns in mind. Interestingly, the literature agrees that fairness interventions should not be standardized and reactive to prior single instances of organizational infrastructure issues, but proactive, based on a thorough understanding of different stakeholders' needs, and accounting for domain-specific and contextual factors.

In regards to the Proposal, it is not yet clear how fairness practices will be effectively operationalized given the mechanisms envisioned in Articles 43 and 61 from the Proposal, respectively for conformance checking and post-market monitoring of high-risk systems. For those systems, providers will be demanded to draft and verify their conformance through a *quality management system*, *technical documentation*, and *post-market monitoring* under the lens of a national body. This body will be guided by a national supervisory authority in coordination with the EDPB (European AI Board from the EU commission). Yet, some detractors, in line with some concerns over organizations' ethics washing, advanced skeptical doubts on the procedural efficacy of these auditing mechanisms [60, 64]. Doubts were related to the undisclosed nature of conformity declarations as well as the nature of contributions of data criteria input to the *EU database for stand-alone high-risk AI systems* in Article 60, withheld from the scrutiny of those affected by such systems and available only upon regulatory bodies' request. This loose gravity towards the public interest might not permit to enforce EU citizen fundamental rights to decide whether a system should be listed as high-risk. In light of the concerns for more structural fairness practices, the evolution of an overly rigid and costly compliance environment could critically undermine these needs. An official impact assessment has been proposed [95] to quantify these costs. Mueller [72] advanced an analysis of the economic costs that could arise for EU small and medium enterprises and corporations. In the forecast, effects will push away venture capital investors, drain European talents and tighten stronger external dependencies leading to a highly unfavorable European environment, with the risk of being excluded from the global AI market. Academics and policy analysts have advanced a debate on the validity of those claims, picturing less-burdening assessments over quality management systems, thus calling the report factitious [37, 57]. Future predictions will need to account both for amendments to the terminology and procedures. Foremost, central analysis focus should be given to the ecosystem of digital technology regulations that the EU has on its agenda [80]. These digital Proposals constitute the European intention of enforcing its legislative sovereignty and set standards for the international market. Leveraging the *Brussels Effect* [12, 31] and the current rise of AI ethics attention across a wide range of institutional and academic stakeholders [35, 102], it is reasonable to predict that in the near future current investments in integrating fairness governance practices could be streamlined into more mature and efficient regulatory frameworks with lower procedural costs while mitigating reputational risks [92].

5 A SOCIOTECHNICAL PERSPECTIVE ON BIAS AND FAIRNESS IN AI

Regarding AI fairness and discrimination, many have pointed out that AI is not merely a tool, it is a sociotechnical endeavour, meaning that the development, use of (and harm from) AI technologies can not be separated from their specific social contexts [27, 90]. When attempting to prevent harm from technologies we must look closely at a new technology's actual capacities and functions within these contexts. An over-emphasis of the role of specific technological features of AI in either causing, or preventing, discrimination, for example, can obscure other forms of discrimination that are occurring, as well as lead to an unproductive and ultimately distracting focus on *fixing* or regulating those specific features [33, 90].

Veale and Borgesius [109] make a similar argument in regards to the Proposal. They cite the examples of the prohibition against releasing AI systems that use subliminal or subconscious techniques to distort a person's behaviour and argue that this focus on evocative, *ripped from the headlines* potential harms does little to mitigate actual harms and adds little to existing legislation [109]. Issues include, for instance, that prohibition only covers manipulative systems that cause individual harm but not a collective harm or a *harm that arises from dynamics of the user-base entwined with an AI system* [109] and that there must be intent to distort behaviour. Dourish and Bell [25] identified a similar phenomenon surrounding the discussion and implementation of ubiquitous computing technologies and contrast the *myth* used to build visions of technologies and the *messiness* of the practical implementation of technologies in reality. They further describe ubiquitous computing researchers as explaining away limitations and unexpected consequences of specific systems by referring to a proximate future where the given technology will be fully realized and highly useful, as soon as a few remaining kinks (such as unevenly distributed infrastructure, for example) are ironed out [25].

In the case of the *messy* realities of AI, it is widely acknowledged that it is non-trivial to build *error-free* models and good quality data within the context of societal factors and power structures at play [18, 27, 67]. To give a specific example, data workers who are frequently manually labeling, cleaning, and enriching the data used for training AI models, have a crucial role in the development of AI systems and their practices are subject to a myriad of non-objective influences [67]. Similarly, the harms often identified with AI use online, such as hyper-personalization, invasion of privacy, and spread of hate speech can stem from issues beyond the technology, such as monopolies, data power imbalances, and un-checked corporate crime [24]. Some have argued that those aspects of online life are a requisite feature of an emerging economic system that has grown out from the existing capitalist economic system [117].

Therefore, we must acknowledge the systemic sources of the discrimination when mitigating discriminatory harm of AI technologies and the discussion of the impact of such technologies should start at an earlier point. In particular, we must look at the specific setting of a given case. This includes considering what specific sociopolitical goals a given AI system is enforcing. For example, in Austria, a risk assessment algorithm created for use in the public employment system has been described as guided by a philosophy of neo-liberal austerity in the social sector which has been replacing the concept of the European welfare state [3]. We must also consider where the discussions are happening, who is involved in the discussions, and how the population is able to discuss and enforce whether an AI in a domain should be used at all. In regards to the Proposal, according to [109], there is evidence of industry influence in high level policy decision-making surrounding the current Proposal.

Another complication in regulating and mitigating harm from AI is the complexity of determining how, or if, it is possible to distinguish between AI decisions and human decisions. If we do not acknowledge these entanglements, there is a risk of bias being addressed with overly mechanistic approaches. In reference to the example of privacy ethics, Nissenbaum [75] has described how a focus on the very attempt to mitigate privacy concerns by ever more sophisticated anonymization methods can lead to overlooking other issues, such as algorithms that do not infringe on privacy, yet are still harmful. Similarly, a focus on attempting to operationalize a very specific concept of fairness, and to regulate specific methods for monitoring it, risks pulling awareness from other algorithmic harms, or even obfuscating underlying causes of harm [7, 90]. In the case of the Austrian AMS, described above, the controversy of a proposed algorithm opened up a whole discussion about how a Public Employment System should be run overall. From the perspective of power aware analysis [67] everyone affected needs to be involved in those decisions.

6 A PHILOSOPHICAL PERSPECTIVE ON BIAS AND FAIRNESS IN AI

We also look at developments in AI and algorithmic fairness through the lens of moral philosophy, specifically normative ethics [49], which essentially investigates the question of whether something is morally right or wrong. There are two major schools of thought in normative ethics; (i) *Deontological ethics* argues the existence and significance of inherent rightness of an action (examples include Kant's *categorical imperative* [86], and Rawls' *veil of ignorance* [93]); (ii) *Consequentialism* judges the morality of an action based on the value it brings (examples include *welfarism* [51], hedonism [70]). While our deontological views inform the building blocks of morality in today's society (e.g., EU fundamental rights), consequential approaches enjoy scalability through the use of representative or proxy metrics in real-world usages (e.g., cost-benefit analysis [58] or per-capita income in economics, and overall accuracy in machine learning as discussed in section 2). Traditional AI research often follows a declarative approach where a mathematical objective is designed and optimized while caring less about the decision-making process and its correctness or representativeness [13, 16, 29]. Such an approach can be argued to be a consequentialist's approach to AI whereby only the optimization of final objective matters and the end justifies the procedure. However, this approach has received a lot of critique within the AI domain, and a range of issues have been pointed out; for example concerning causality [17, 34], fairness [29, 66], explainability [13], including the comparability and robustness of explanations [88, 89], and trustworthiness [107].

Another angle from which AI developments can be looked at, is *Welfarism* [51] (a type of consequentialism), which suggests choosing the action that maximizes the welfare or well-being of the population. In fact, it is widely used in some areas of economics, game theory, social-choice theory, and applications. Welfarism is often studied in two major forms; (i) *Utilitarianism* [104] emphasizes maximizing the welfare of the population; (ii) *Egalitarianism* argues for equality often leading to a form of Rawlsian justice [93] which comes under deontological ethics, but its objective form in welfarism tries to maximize the welfare of the worst-off. Utilitarianism is found to be heavily embedded in today's society. For example, the optimization objectives (loss functions) in machine learning are often the aggregate errors over the set of data points or the individuals, i.e., utilitarian in nature. Utilitarian social welfare is quite prevalent in economics, computational social choice (allocation, voting, etc.)[1]. Such utilitarian objectives tend to optimize for the overall utility while may be best-serving the majority and poorly serving minority populations. This is one of the reasons due to which the usual loss-minimizing objectives have been found to be unfair in many applications including criminal justice, banking, and gig-economy. On the other hand, egalitarian welfarism in machine learning would likely try to equalize the errors of all or groups of individuals instead of minimizing the aggregate errors. In fact algorithmic fairness notions like individual fairness [26], equal opportunity and equality of odds [40], equal mistreatment [114] are either inspired by or promote egalitarian views in consequential modeling (error represents a consequence). These notions have been found to reduce the effects of pre-existing biases in data and to improve the utilities of marginalized groups under algorithmic decision-making systems.

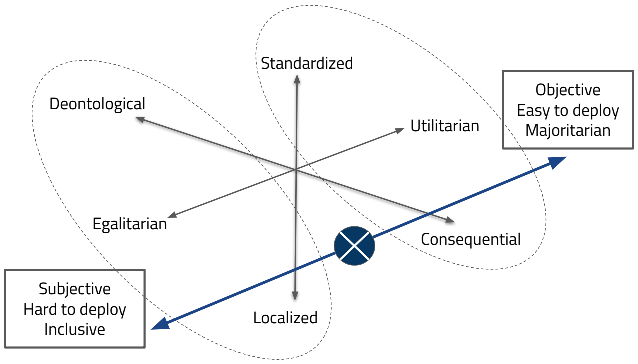

Figure 1: Three identified axes along which the debate about regulation of fairness in AI falls; Here they are aligned along high-level characterizations and common arguments made for, and against, each of the ends of the identified axes.

A few recent works have also explored non-consequential or deontological approaches to algorithmic fairness. These works can be grouped into two categories. (1) Works on *procedural fairness* [30, 32] argue that it is essential for the chosen design and principles to be socially acceptable. Thus, these works focus on understanding how people assess fairness and ways to infer societal expectations about fairness principles thereby accounting for all voices in designing decision-making systems. For example, Grgić-Hlača et al. [32] propose a framework for procedural fairness by evaluating the moral judgments of humans regarding the use of certain features and accordingly designing decision-making systems. (2) Another set of works argue for causal and counterfactual fairness, i.e., addressing unfair causal effects of sensitive attributes in the decision-making process [17, 55]. Instead of focusing on the outcome alone, these works have explored deontological aspects and propose to ensure fairness in the decision-making process.

7 MAPPING PERSPECTIVES

We have identified three axes along which different perspectives in the debate about AI Regulation for preventing unfairness and discrimination fall. These axes may or may not be orthogonal, i.e., they may have relationships with each other. In the following sections, we define the axes and describe the debate surrounding regulating fairness in AI represented by each axis. These are not all of the axes of debate, rather these are salient tensions that we have

[1] Nash social welfare [50] is an exception.

identified. We find them helpful in conceptualizing and mapping the values and desiderata of the perspectives we are focusing on.

7.1 Axis-1: Standardization vs. Localization

7.1.1 **The axis:** This first axis of interest addresses the differences between standardization and localization.

Standardization entails *the process of making things of the same type all have the same basic features* (from Cambridge dictionary), specifically, through the creation of protocols to guide the design, development, and creation of such goods or services based on the consensus of all the relevant parties in the industry. Standardization is intended to ensure that all the goods and services produced respecting those protocols come with the same or equivalent quality, safety, interoperability and compatibility. For this reason, multiple parties need to be involved in developing such protocols and standards, namely, manufacturers, sellers, buyers, customers, trade associations, users or regulators (https://www.iso.org/standards.html). By contrast, localization describes *the process of making a product or a service more suitable for a particular country, area, etc.* (from Cambridge dictionary). In essence, localization entails adapting the product or service to the characteristics of a given culture, region, or society.

7.1.2 **Pros and cons:** In the context of AI, advocates for and members of industry frequently cite standardization as a method for preventing or mitigating discrimination [41, 53, 108]. In this respect, high-risk AI systems will be presumed to comply with the requirements established in the AI Proposal if they are, as well, in conformity with the harmonised standards published by the Official Journal of the European Union as referred to in article 40 [85]. Likewise, high-risk AI systems in conformity with the specifications referred to in Article 41 of the AI Proposal will be presumed in conformity with the regulation [85]. In this sense, conformity with standards and specifications as proposed in the AI Regulation will allow the entry of high-risk AI systems in the European market while guaranteeing agreed levels of quality and safety that ensure the adherence to European principles and values (i.e., non-discrimination, fairness, and human dignity).

A dilemma regarding standardization, however, appears when there is a disagreement regarding the standard of fairness that should be used to assess AI systems. As presented in section 2.1 the straightforward example of incompatible fairness standards referred to the case of COMPAS and the different standards followed by ProPublica [5] and Northpoint [22] for their fairness assessments, i.e., disparate mistreatment and calibration respectively [118]. Moreover, overly specific and strict standards and frameworks risk encoding a biased, restrictive, non-relevant to everyone, singular worldview, and may ultimately lead to uniformization from a top-down approach section 4. In truth, standardarization as a method to enforce fairness can in some cases overlook the root-causes of bias, setting standards and notions of fairness that do not offer a real solution to the intrinsic discrimination or biases in certain situations or contexts section 5. A—purely hypothetical—example of this problem would be the hard-coded requirements for gender parity in school admissions or hiring where there was a low representation of one of the genders, e.g., due to relocation for work reasons or armed conflicts. The solution would be to establish an acceptable ratio of males to females set at a level appropriate to the local context, rather than a strict gender parity requirement.

In this regard, localizing AI systems entails the process of making them local in character by limiting the ethics regulation and specifics of enforcement to the desired area. Whereas the complete localization of AI systems will be in conflict with the embedded values of the AI Regulation (e.g., European Common Market and European Fundamental Rights), the localization of some of the decisions regarding their design, development, or deployment may allow a more tailored approach to address AI discrimination and biases in specific geographical, cultural, or sociotechnical contexts. The localization of some requirements and technical solutions may, as well, allow for the definition of ethical and legal guidelines that address the specific circumstances of a community, local area, or sector beyond the general standards and specifications.

7.2 Axis-2: Utilitarian vs. Egalitarian

7.2.1 **The axis:** The second axis of interest addresses differences between utilitarian and egalitarian views. While a utilitarian philosophy is one of maximizing the overall welfare of the population, egalitarianism aims for equality amongst all those people.

7.2.2 **Pros and cons:** Utilitarianism has long been argued to be in conflict with the certain conceptualizations of fairness (see Chapter 14 of Hooker [45]). In the context of AI, algorithms are often designed to optimize for certain mathematical objectives (which can be categorized as a declarative approach). The objective functions in machine learning tasks usually measure a form of aggregate accuracy over a population, which fits the definition of a utilitarian measure. Optimizing solely for such a measure in AI applications risks optimizing the utility of the whole population while hurting minority groups in many [40, 114]. Utilitarian approaches are so ingrained in the computing research and development mindset that the early group fairness notions—which are supposed to mitigate the discriminatory effects of utilitarian objectives—such as demographic parity, had been reduced to utilitarian forms by constraining over the aggregate benefits or outcomes of groups of individuals [115]. The literature has now moved on to notions such as individual fairness, equal opportunity, and treatment parity which, even though outcome-based, are more egalitarian in nature.

Despite its obvious conflicts with fairness, and egalitarianism's close connection with fairness, utilitarian welfare is often cited a necessary factor in system and policy design. In fact, protecting the EU's economic interests is stated as a goal of the AI Act [85]. Since utilitarianism captures a certain overall efficiency of a system (accuracy in machine learning, utilitarian welfare in economics), its goals often reflect business-oriented metrics of AI applications (i.e., click-through rate for recommendations in online marketplaces, or success-rate of ranked workers on gig-economy platforms). However, there might be a trade-off between maximizing efficiency and achieving other social objectives like equity or fairness in cases of inherent imbalance in the data or population [9, 10].

7.3 Axis-3: Consequential vs. Deontological

7.3.1 **The axis:** This third axis of interest from the discussions in sections 3 to 6 represents the differences between consequential and deontological ethics. Deontological ethics argue for the existence of the inherent rightness of an action, while consequential ethics evaluate morality based on the consequences of an action.

7.3.2 **Pros and cons:** Technical measures for mitigating AI based discrimination tend to focus on fairness notions, whereby a fairness constraint is often added to the original objective. Fairness in this case is defined by statistical properties of the outcome/decision of the system (e.g., demographic parity). Fairness notions thus seek to reduce harm by adjusting or influencing the outcome to fit some statistical definition of fairness. While the motivation for doing this may be based on deontological principles of equality, this approach belies a consequentialist definition of fairness, wherein one declares that fairness has been achieved through an equality in outcome, such as equal amount of good (accurate) and bad (inaccurate) outcomes for each group.

Deontological ethics is often given as an opposite to consequentialism. A deontological approach argues for the existence and significance of the inherent rightness of an action; in the context of AI based discrimination, this would suggest that the approach described above does not meet the criteria of acting morally, as the focus is on shifting the outcome. From a deontological perspective, an AI system is unlikely to be fair if the development of AI itself is not driven by essential guiding principles, such as fairness.

The Proposal's prohibition of certain uses is based on deontological principles of protecting fundamental individual rights. However, the risk based approach could be viewed as consequential, in that it only targets systems used in contexts perceived as being highly consequential. This means that many AI systems which might exhibit harmful representational or discriminatory biases, such as social media and online platforms are relieved of any requirements.

Summary: Based on the pattern of high-level characterizations and common arguments made for, and against, each end of the identified axes, we place them along a single axis, with one end containing localized, deontological, egalitarian approaches (LED) and the other end containing standardized, utilitarian, consequential approaches (SUC); we illustrate this mapping in Figure 1. The LED end contains approaches that purport to acknowledge systemic and complex causes of discrimination and are often criticized as being overly subjective and hard to deploy. The approaches on the SUC end purport to be objective and easy to implement while often being critiqued as failing to recognize systemic causes or ensure inclusion of minority voices. This mapping of the perceived benefits and shortcomings of each approach allows us to identify a key tension in the debate on regulating fairness in AI. It is one that is based on differing understandings of the nature of bias and discrimination, along with differing priorities as to what constitutes practicality and implementability in efforts to increase fairness. Following this, we suggest how the Proposal could better balance these values, as well as the differing perspectives of stakeholders, to achieve the stated goal of guaranteeing agreed levels of quality and safety in accordance with European principles and values (i.e., non-discrimination, fairness, and human dignity) without creating major hurdles for the European AI Industry.

8 KEY AGREEMENT AND A PATH FORWARD

8.1 Key Agreement

We see a specific agreement amongst the presented perspectives, regarding limitations of the current regulation. Ultimately each of the perspectives agree that regulation needs to be grounded in the reality of the context of the use of AI, and that this is not sufficiently achieved in the Proposal. A brief summary of these previously discussed *realities* that the Proposal as not sufficiently accounting for is as follows: (1) lack of agreement on what technology like AI really *is* and what are its capabilities, (2) cost and complexity for a business to follow the required regulations, (3) the known limitations of debiasing techniques and explanations of black boxes, (4) lack of specifications on how to best implement *human oversight* in the context of AI systems, (5) varied and shifting notions of fairness within society, (6) impact of power imbalances (eg. technological divide, data power, company size, and market share) on the creation and enforcement of and ability to comply with the Proposal.

8.2 A Path Forward: Balancing Perspectives

8.2.1 **Standardization and Localization.** Standardization may facilitate the translation of fundamental rights, i.e., right to fairness, into standards and specifications to be followed and complied with by all AI actors with the aim of ensuring that AI systems do not discriminate nor mistreat individuals.

Likewise, localization may allow the clarification of deontological values in more specific and concrete requirements, metrics, or assessments, particular to each enforcement context. This is to prevent a top-down enforcement of operationalizations of fairness that are untenable, or even unfair, in some contexts. For example, in section 4 we have summarized the literature demonstrating that ensuring fairness compliance from AI industry could as well be served from a more localized approach to operationalizing fairness. This does not imply the relativization of the legal and ethical principle of fairness but, on the contrary, take into account the wider scenario beyond the purely technical nature of AI and strengthen the enforcement of fairness during the whole life cycle of AI.

Proposed role of AI Regulation. Standardization should be used to the extent that the measure has a direct link to upholding the deontological value of fairness. In order to ensure the principle of universalization, though, special care must be taken to build in flexible localization allowances.

8.2.2 **Utilitarian and Egalitarian.** It may be possible to maintain an egalitarian approach to AI Regulations, while also taking advantage of the potential benefits of utilitarian measures. For example, to promote equality (i.e., bring in egalitarianism) all stakeholders could be given sufficient power to provide inputs on how to maximize and measure their welfare. Any decisions about utilitarian measures would then be based on this input. Note that increased awareness of the use of AI systems and their implications toward fairness among the responding individuals (stakeholders) is essential for a successful process. This approach would, again, bring up the question of standardization versus localization. Specifically, how highly localized measures would be required to adequately account for the policy expectations of all individuals in an egalitarian fashion. To address this, we would defer to the principles suggested in section 8.2.1. Extensive work is needed to determine how best to implement such a process, but some of the open questions may be best left answered by the inclusive input process itself.

Proposed role of AI Regulation. The specific framework for how to obtain and incorporate stakeholder inputs should be laid out. A way needs to be found to enforce that *all* stakeholders have sufficient power and influence in AI Regulation decision making processes and that they are themselves sufficiently aware of the potential adverse implications of AI technology.

8.2.3 **Deontological and Consequential.** The EU's stance on fairness is deontological, in that fairness is justified by itself, with no direct subordination to its eventual outcomes. What matters is whether the action is motivated by duty (respect of the *moral law*: dignity and universalization). However, expectations of individuals on the specifics of what constitutes freedom, equality, and dignity, may vary across cultures, geographies, and contexts. This has led digital and human rights groups to highlight that AI policies should empower individuals, communities, and organisations to contest AI-based systems and to demand redress when they themselves determine that their fundamental rights have been violated [7].

The Proposal itself is not intended to legislate individual rights; that is intended to be covered in other laws of the European legal framework. With that in mind, the Proposal could still enforce an individual's need to be informed and to understand the impacts. Therefore transparency, explainability of the design, development and implementaion of AI systems, as well as their output, remains paramount. There must also be understandable and effective methods for stakeholders to adjust the specific standards, such as what uses are forbidden, in the case of unforeseen use cases and impacts or of the recognition of previously ignored violations of the European principles.

Proposed role of AI Regulation. Requirements such as documentation and transparency should specifically serve stakeholders' needs to understand the implications of AI systems for their specific situation, life, and work.

9 CONCLUSION

In this position paper, we presented technical, legal, industrial, sociotechnical, and (moral) philosophical perspectives on the debate on fairness in AI systems with a particular focus on the Proposal of the EU AI Act. We identified a pattern of common arguments representing a key tension in the debate with one side containing *deontological, egalitarian, localized* approaches and the other side containing *standardized, utilitarian, consequential* approaches. We discussed how different (symbolic) ends of the axes could be reconciled and proposed the following roles that the AI Regulation could take to successfully address these tensions: **(1)** apply standardization to uphold deontological values, but ensure universalization by including flexible localization allowances; **(2)** lay out a framework to incorporate stakeholder inputs and ensure that they are sufficiently aware of potential adverse implications of AI technology; and **(3)** design requirements of documentation and transparency so that they serve the needs of stakeholders.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS - Artificial Intelligence without Bias" and (grant agreement number 860621) for the project "NL4XAI - Natural Language for Explainable AI". The authors would like to thank all the panelists of European AI Regulation Week 2021 (https://aia-week.pages.citius.usc.es/) for their helpful comments on AI regulation. This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] Angelika Adensamer and Lukas Daniel Klausner. 2021. "Part Man, Part Machine, All Cop": Automation in Policing. *Frontiers in Artificial Intelligence* 4 (2021), 29. https://doi.org/10.3389/frai.2021.655486

[2] Edward Tristram Albert. 2019. AI in talent acquisition: A review of AI-applications used in recruitment and selection. *Strategic HR Review* 18, 5 (2019), 215–221. https://doi.org/10.1108/shr-04-2019-0024

[3] Doris Allhutter, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. 2020. Algorithmic profiling of Job Seekers in austria: How austerity politics are made effective. *Frontiers in Big Data* 3 (2020). https://doi.org/10.3389/fdata.2020.00005

[4] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20, 3 (2018), 973–989. https://doi.org/10.1177/1461444816676645

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine Bias*. *Ethics of Data and Analytics* (2022), 254–264. https://doi.org/10.1201/9781003278290-37

[6] Jacqui Ayling and Adriane Chapman. 2021. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics* (2021), 1–25. https://doi.org/10.1007/s43681-021-00084-x

[7] Agathe Balayan and Seda Gürses. 2021. *Beyond Debiasing: Regulating AI and Its Inequalities.* Technical Report. Delft University of Technology.

[8] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. 2019. *Consumer-lending discrimination in the FinTech era.* Technical Report. National Bureau of Economic Research.

[9] Richard Berk et al. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

[10] Dimitris Bertsimas, Vivek F. Farias, and Nikolaos Trichakis. 2012. On the efficiency-fairness trade-off. *Management Science* 58, 12 (2012), 2234–2250. https://doi.org/10.1287/mnsc.1120.1549

[11] Miranda Bogen and Aaron Rieke. 2018. *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias.* Report. Upturn.

[12] Anu Bradford. 2020. *The Brussels effect: How the European Union rules the world.* Oxford University Press, USA.

[13] Nadia Burkart and Marco F. Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 70 (2021), 245–317. https://doi.org/10.1613/jair.1.12228

[14] 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://doi.org/10.1126/science.aal4230

[15] Longbing Cao. 2022. AI in finance: Challenges, techniques, and opportunities. *Comput. Surveys* 55, 3 (2022), 1–38. https://doi.org/10.1145/3502289

[16] Manuel Carabantes. 2020. Black-box artificial intelligence: an epistemological and critical analysis. *AI & SOCIETY* 35, 2 (2020), 309–317. https://doi.org/10.1007/s00146-019-00888-w

[17] Daniel C Castro, Ian Walker, and Ben Glocker. 2020. Causality matters in medical imaging. *Nature Communications* 11, 1 (2020), 1–10. https://doi.org/10.1038/s41467-020-17478-w

[18] Kyla Chasalow and Karen Levy. 2021. Representativeness in Statistics, Politics, and Machine Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* Association for Computing Machinery, New York, NY, USA, 77–89. https://doi.org/10.1145/3442188.3445872

[19] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. 2019. Translation, tracks & data: an algorithmic bias effort in practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–8. https://doi.org/10.1145/3290607.3299057

[20] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David 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://doi.org/10.1145/3351095.3372878

[21] Abhisek Dash, Abhijnan Chakraborty, Saptarshi Ghosh, Animesh Mukherjee, and Krishna P Gummadi. 2021. When the umpire is also a player: Bias in private label product recommendations on e-commerce marketplaces. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 873–884. https://doi.org/10.1145/3442188.3445944

[22] William Dieterich, Christina Mendoza, and MS Tim Brennan. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity.

[23] Tom Dobber, Ronan Ó Fathaigh, and Frederik Zuiderveen Borgesius. 2019. The regulation of online political micro-targeting in Europe. *Internet Policy Review* 8, 4 (2019).

[24] Cory Doctorow. 2021. *How to Destroy 'Surveillance Capitalism'.* Medium Editions.

[25] Paul Dourish and Genevieve Bell. 2011. *Divining a Digital Future: Mess and Mythology in Ubiquitous Computing.* MIT Press, Cambridge, Mass.

[26] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12.* https://doi.org/10.1145/2090236.2090255

[27] M. C. Elish and danah boyd. 2017. Situating methods in the magic of Big Data and ai. *Communication Monographs* 85, 1 (2017), 57–80. https://doi.org/10.1080/03637751.2017.1375130

[28] Robert Epstein. 2019. Why Google Poses a Serious Threat to Democracy, and How to End That Threat. *America Institute for Behavioral Research and Technology* (2019).

[29] Jessie Finocchiaro, Roland Maio, Faidra Monachou, Gourab K Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis. 2021. Bridging Machine Learning and mechanism design towards Algorithmic Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* https://doi.org/10.1145/3442188.3445912

[30] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency.* 90–99. https://doi.org/10.1145/3287560.3287563

[31] Graham Greenleaf. 2021. The 'Brussels Effect' of the EU's 'AI Act' on Data Privacy Outside Europe. , 3-7 pages. https://papers.ssrn.com/abstract=3898904

[32] Nina Grgić-Hlača, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference - WWW '18.* 903–912. https://doi.org/10.1145/3178876.3186138

[33] 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 *Thirty-Second AAAI Conference on Artificial Intelligence.* https://ojs.aaai.org/index.php/AAAI/article/view/11296

[34] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)* 53, 4 (2020), 1–37. https://doi.org/10.1145/3397269

[35] Abhishek Gupta, Connor Wright, Marianna Bergamaschi Ganapini, Masa Sweidan, and Renjie Butalid. 2022. State of AI Ethics Report (Volume 6, February 2022). *arXiv preprint arXiv:2202.07435* (2022).

[36] Alisha Haridasani Gupta. 2019. Are Algorithms Sexist? *The New York Times* (2019).

[37] Meeri Haataja and Joanna J. Bryson. 2021. *What costs should we expect from the EU's AI Act?* SocArXiv. Center for Open Science.

[38] Philipp Hacker. 2018. Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review* 55, 4 (2018), 1143–1185. https://doi.org/10.54648/cola2018095

[39] Bernard E. Harcourt. 2007. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age.* University of Chicago Press. viii, 336 pages.

[40] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, Vol. 29. 3315–3323. https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[41] John C. Havens. 2018. Creating the human standard for ethical autonomous and intelligent systems (A/IS). *AI Matters* 4 (4 2018), 28–31. Issue 1. https://doi.org/10.1145/3203247.3203255

[42] Thomas T Hills. 2019. The Dark Side of Information Proliferation. *Perspectives on Psychological Science* 14 (2019), 323–330. https://doi.org/10.1177/1745691618803647

[43] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. 22, 7 (06 2019), 900–915. https://doi.org/10.1080/1369118x.2019.1573912

[44] 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://doi.org/10.1145/3290605.3300830

[45] Brad Hooker. 2014. *Utilitarianism and fairness.* Cambridge University Press, 280–302.

[46] Dietmar Hübner. 2021. Two kinds of discrimination in AI-based penal decision-making. *ACM SIGKDD Explorations Newsletter* 23, 1 (2021), 4–13. https://doi.org/10.1145/3468507.3468510

[47] Javier Camacho Ibáñez and Mónica Villas Olmeda. 2021. Operationalising AI ethics: How are companies bridging the gap between practice and principles? An exploratory study. (08 2021). https://doi.org/10.1007/s00146-021-01267-0

[48] Information Commissioner's Office (ICO). 2021. Guide to the General Data Protection Regulation (GDPR). https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/

[49] Shelly Kagan. 2018. *Normative ethics.* Routledge.

[50] Mamoru Kaneko and Kenjiro Nakamura. 1979. The Nash Social Welfare function. *Econometrica: Journal of the Econometric Society* 47, 2 (1979), 423–435. https://doi.org/10.2307/1914191

[51] Simon Keller. 2009. Welfarism. *Philosophy Compass* 4, 1 (2009), 82–95. https://doi.org/10.1111/j.1747-9991.2008.00196.x

[52] 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, 1–23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[53] Ansgar Koene, Adam Leon Smith, Takashi Egawa, Sukanya Mandalh, and Yohko Hatada. 2018. IEEE P70xx, Establishing Standards for Ethical Technology. *Proceedings of KDD, ExCeL London UK* (8 2018), 1–2.

[54] P. M. Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 2019. Defining AI in Policy versus Practice. https://papers.ssrn.com/abstract=3431304

[55] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, Vol. 30. 4066–4076. https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[56] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science* 65, 7 (2019), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093

[57] Moritz Laurer, Andrea Renda, and Timothy Yeung. 2021. *Clarifying the costs for the EU's AI Act.* Technical Report.

[58] Richard Layard and Stephen Gllaister. 1994. *Cost-benefit analysis.* Cambridge University Press, Cambridge, UK.

[59] Michelle Seng Ah Lee and Jatinder Singh. 2021. The landscape and gaps in open source fairness toolkits. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3411764.3445261

[60] Mark MacCarthy and Kenneth Propp. 2021. Machines learn that Brussels writes the rules: The EU's new AI regulation. *Brookings, May* 4 (2021), 2021.

[61] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–14. https://doi.org/10.1145/3313831.3376445

[62] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. On the applicability of machine learning fairness notions. *ACM SIGKDD Explorations Newsletter* 23, 1 (2021), 14–23. https://doi.org/10.1145/3468507.3468511

[63] Karl Manheim and Lyric Kaplan. 2019. Artificial intelligence: Risks to privacy and democracy. *Yale JL & Tech.* 21 (2019), 106. https://ssrn.com/abstract=3273016

[64] Ian Manners. 2002. Normative Power Europe: A Contradiction in Terms? 40, 2 (06 2002), 235–258.

[65] Mark McFadden, Kate Jones, Emily Taylor, and Georgia Osborn. 2021. Harmonising Artificial Intelligence: The Role of Standards in the EU AI Regulation. (2021).

[66] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (jul 2021), 35 pages. https://doi-org.ezbusc.usc.gal/10.1145/3457607

[67] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–25. https://doi.org/10.1145/3415186

[68] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. 8, 1 (03 2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902 arXiv:1811.07867

[69] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11 (11 2019), 501–507. https://doi.org/10.1038/s42256-019-0114-4

[70] Andrew Moore. 2013. Hedonism. Stanford University. https://plato.stanford.edu/entries/hedonism/

[71] Jessica Morley, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mökander, and Luciano Floridi. 2021. Ethics as a Service: A Pragmatic Operationalisation

of AI Ethics. 31, 2 (2021), 239–256. https://doi.org/10.1007/s11023-021-09563-w

[72] Benjamin Mueller. 2021. *How Much Will the Artificial Intelligence Act Cost Europe?* Technical Report. Center for Data Innovation.

[73] Sendhil Mullainathan. 2018. Algorithmic fairness and the social welfare function. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 1–1. https://doi.org/10.1145/3219166.3219236

[74] Catelijne Muller. 2020. *The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law*. Technical Report. Council of Europe, Strasbourg.

[75] Helen Nissenbaum. 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press. https://doi.org/10.1515/9780804772891

[76] Eirini Ntoutsi et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10 (5 2020). Issue 3. https://doi.org/10.1002/widm.1356

[77] Council of the European Union. 2000. Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation.

[78] Council of the European Union. 2000. Council Directive E 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin.

[79] Council of the European Union. 2004. Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services.

[80] Commission of the European Union. 2021. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions 2013 Digital Compass: the European way for the Digital Decade.

[81] Will Orr and Jenny L Davis. 2020. Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society* 23, 5 (2020), 719–735. https://doi.org/10.1080/1369118x.2020.1713842

[82] European Parliament and Council. 2007. Charter of Fundamental Rights of the European Union.

[83] European Parliament and Council of the European Union. 2006. Directive 2006/54/EC Of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation.

[84] European Parliament and Council of the European Union. 2016. Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[85] European Parliament and Council of the European Union. 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.

[86] Herbert James Paton. 1971. *The categorical imperative: A study in Kant's moral philosophy*. Vol. 1023. University of Pennsylvania Press.

[87] 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. https://doi.org/10.1145/3366423.3380196

[88] Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. 2021. CARLA: A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 2021*.

[89] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. On Counterfactual Explanations under Predictive Multiplicity. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020 (Proceedings of Machine Learning Research, Vol. 124)*. AUAI Press, 809–818.

[90] Seeta Peña Gangadharan and Jędrzej Niklas. 2019. Decentering Technology in Discourse on Discrimination. *Information, Communication & Society* 22, 7 (June 2019), 882–899. https://doi.org/10.1080/1369118X.2019.1593484

[91] Inioluwa Deborah Raji et al. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44. https://doi.org/10.1145/3351095.3372873

[92] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23. https://doi.org/10.1145/3449081

[93] John Rawls. 2009. *A theory of justice*. Harvard University Press, Cambridge, Mass.

[94] Richard M. Re and Alicia Solow-Niederman. 2019. Developing artificially intelligent justice. *Stan. Tech. L. Rev.* 22 (2019), 242. https://ssrn.com/abstract=3390854

[95] Andrea Renda et al. 2021. *Study to support an impact assessment of regulatory requirements for artificial intelligence in Europe*. Technical Report. European Commission - Directorate-General for Communications Networks, Content and Technology.

[96] Filipe N. Ribeiro et al. 2018. Media Bias Monitor : Quantifying Biases of Social Media News Outlets at Large-Scale. In *Twelfth International AAAI Conference on Web and Social Media*. AAAI Press, Palo Alto, California, 290–299. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17878

[97] Brianna Richardson and Juan E. Gilbert. 2021. A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions. (12 2021). arXiv:2112.05700 http://arxiv.org/abs/2112.05700

[98] Boris Ruf and Marcin Detyniecki. 2021. Towards the Right Kind of Fairness in AI. (09 2021). arXiv:2102.08453 [cs] http://arxiv.org/abs/2102.08453

[99] Ajay Sandhu and Peter Fussey. 2021. The 'uberization of policing'? How police negotiate and operationalise predictive policing technology. *Policing and Society* 31, 1 (2021), 66–81. https://doi.org/10.1080/10439463.2020.1803315

[100] Cristian Santesteban and Shayne Longpre. 2020. How big data confers market power to Big Tech: Leveraging the perspective of data science. *The Antitrust Bulletin* 65, 3 (2020), 459–485. https://doi.org/10.1177/0003603x20934212

[101] Laura Sartori and Andreas Theodorou. 2022. A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. *Ethics and Information Technology* 24, 1 (2022), 1–11. https://doi.org/10.1007/s10676-022-09624-3

[102] Daniel S. Schiff, Kelly Laas, Justin B. Biddle, and Jason Borenstein. 2022. Global AI Ethics Documents: What They Reveal About Motivations, Practices, and Policies. In *Codes of Ethics and Ethical Guidelines: Emerging Technologies, Changing Fields*. Springer International Publishing, 121–143. https://doi.org/10.1007/978-3-030-86201-5_7

[103] 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* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[104] Amartya Sen. 1979. Utilitarianism and welfarism. *The journal of Philosophy* 76, 9 (1979), 463–489. https://doi.org/10.2307/2025934

[105] Josh Simons and Dipayan Ghosh. 2022. *Utilities for democracy: Why and how the Algorithmic Infrastructure of Facebook and Google must be regulated*. https://www.brookings.edu/research/utilities-for-democracy-why-and-how-the-algorithmic-infrastructure-of-facebook-and-google-must-be-regulated/

[106] Taylor Telford. 2019. *Apple Card algorithm sparks gender bias allegations against Goldman Sachs*. https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/

[107] Ehsan Toreini et al. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 272–283. https://doi.org/10.1145/3351095.3372834

[108] Antje Von Ungern-Sternberg et al. 2022. Discriminatory AI and the Law– Legal standards for algorithmic profiling. In *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives (Cambridge Law Handbooks)*. Cambridge University Press. https://ssrn.com/abstract=3876657

[109] Michael Veale and Frederik Zuiderveen Borgesius. 2021. Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach. *Computer Law Review International* 22, 4 (Aug. 2021), 97–112. https://doi.org/doi:10.9785/cri-2021-220402

[110] 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. https://doi.org/10.1145/3173574.3174014

[111] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness* (New York, NY, USA) *(FairWare '18)*. Association for Computing Machinery, 1–7. https://doi.org/10.1145/3194770.3194776

[112] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. 2021. *On the Fairness of Causal Algorithmic Recourse*. arXiv:2010.06529 [cs, stat] http://arxiv.org/abs/2010.06529

[113] Raphaële Xenidis and Linda Senden. 2019. EU non-discrimination law in the era of artificial intelligence: Mapping the challenges of algorithmic discrimination. In *Ulf Bernitz et al (eds), General Principles of EU law and the EU Digital Order*. Kluwer Law International, 2020, 151–182. https://ssrn.com/abstract=3529524

[114] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180. https://doi.org/10.1145/3038912.3052660

[115] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. 2017. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*. 229–239.

[116] Aleš Završnik. 2020. Criminal justice, artificial intelligence systems, and human rights. In *ERA Forum*, Vol. 20. Springer, 567–583. https://doi.org/10.1007/s12027-020-00602-0

[117] Shoshana Zuboff. 2019. Surveillance Capitalism and the Challenge of Collective Action. *New Labor Forum* 28, 1 (Jan. 2019), 10–29. https://doi.org/10.1177/1095796018819461

[118] Frederik Zuiderveen Borgesius. 2018. *Discrimination, artificial intelligence, and algorithmic decision-making*. Technical Report. Strasbourg Council of Europe.