Ethics and Information Technology (2024) 26:31
https://doi.org/10.1007/s10676-024-09746-w

ORIGINAL PAPER

Policy advice and best practices on bias and fairness in AI

Jose M. Alvarez[1,2] · Alejandra Bringas Colmenarejo[3] · Alaa Elobaid[4,5] · Simone Fabbrizzi[4,6,7] · Miriam Fahimi[8] · Antonio Ferrara[9,10] · Siamak Ghodsi[5,6] · Carlos Mougan[3] · Ioanna Papageorgiou[6] · Paula Reyero[11] · Mayra Russo[6] · Kristen M. Scott[12] · Laura State[1,2] · Xuan Zhao[13] · Salvatore Ruggieri[2]

Accepted: 16 January 2024

Abstract
The literature addressing bias and fairness in AI models (*fair-AI*) is growing at a fast pace, making it difficult for novel researchers and practitioners to have a bird's-eye view picture of the field. In particular, many policy initiatives, standards, and best practices in fair-AI have been proposed for setting principles, procedures, and knowledge bases to guide and operationalize the management of bias and fairness. The first objective of this paper is to concisely survey the state-of-the-art of fair-AI methods and resources, and the main policies on bias in AI, with the aim of providing such a bird's-eye guidance for both researchers and practitioners. The second objective of the paper is to contribute to the policy advice and best practices state-of-the-art by leveraging from the results of the NoBIAS research project. We present and discuss a few relevant topics organized around the NoBIAS architecture, which is made up of a Legal Layer, focusing on the European Union context, and a Bias Management Layer, focusing on understanding, mitigating, and accounting for bias.

Keywords Artificial Intelligence · Bias · Fairness · Policy advice · Best practices

Introduction

The last decade has witnessed a renaissance of Artificial Intelligence (AI), leading to an increasingly pervasive usage in many socially sensitive tasks. However, many concerns have been raised about the—intentional or unintentional—negative impacts on individuals and society due to biases embedded in AI models[1] (Future of Privacy Forum, 2017; Shelby et al., 2023). A few AI incident databases report collections of harms or near harms realized in the real world by intelligent systems (Turri & Dzombak, 2023), the most relevant one being illegal discrimination against social groups protected by non-discrimination law (Altman, 2020). In fact, there is a deep academic and social discussion around the need to evaluate the claims, decisions, actions and policies that are being made based on the AI's alleged neutrality as more examples confirm that algorithmic systems "are value-laden in that they (1) create moral consequences, (2) reinforce or undercut ethical principles, or (3) enable or diminish stakeholder rights and dignity" (Martin, 2019).

The objective of this paper is twofold.

First, we aim at providing the reader with an up-to-date entry-point to the state-of-the-art of the multidisciplinary research on bias and fairness in AI. We take a bird's-eye view of the methods and resources, with links to specialized surveys, and of the issues and challenges related to policies on bias and fairness in AI. Such an overview provides guidance for both new researchers and AI practitioners that want to find their way in the blooming literature of the area.

Second, we contribute towards the objective of providing policy advice and best practices for dealing with bias and fairness in AI by leveraging from the results of the NoBIAS project[2]. We present and discuss a few topics that emerged during the execution of the project, whose focus was on legal challenges in the context of the European Union (EU) legislation, and on understanding, mitigating, and accounting for bias from a multidisciplinary perspective. The presented

[1] Due to the large body of literature, we prioritize the citation of survey papers, where applicable, and recent works.

[2] https://nobias-project.eu/.

Extended author information available on the last page of the article

Published online: 29 April 2024

issues are relevant but not sufficiently developed or acknowledged in the literature. As such, the paper can contribute to the advancement of the research and to increase awareness on bias and fairness in AI.

Introducing fair-AI

In general, bias can be defined as "an attitude that always favors one way of feeling or acting over any other" (Bias, 2023). In human cognition and reasoning, this is the result of evolution (Haselton et al., 2005), for which some heuristics work well in most circumstances, or have a smaller cost than alternative strategies. In AI, biases can originate in the data (*pre-existing bias*), in the design of AI algorithms and systems (*technical bias*), and in the organizational processes using AI models (*emerging bias*). Most AI models are data-driven, hence they may inherit bias embedded in representations of reality encoded in raw data (Shahbazi et al., 2023). In fact, data are not neutral but are instead value-laden (Gitelman, 2013). Biases in AI algorithms have similar foundations as human cognitive biases, namely the reliance on heuristic algorithmic-search strategies that work well on average (Hellström et al., 2020). Quantitative loss metrics that are optimized by AI algorithms may result in an oversimplification of the complexity of reality, hence leading to a systematic difference between what AI actually models and the reality it is intended to abstract (Grimes & Schulz, 2002; Danks and London, 2017) (*internal validity*). Moreover, the usage of AI in complex socio-technical processes under untested or unplanned conditions may suffer from a lack of generalizability of the AI models (*external validity*). Several categorizations of the sources of bias and fairness in AI have been proposed in contexts such as social data (Olteanu et al., 2019), Machine Learning (ML) representations (Shahbazi et al., 2023), ML algorithms (Mehrabi et al., 2021), recommender systems (Chen et al., 2023a), algorithmic hiring (Fabris et al., 2023), large language models (Gallegos et al., 2023), and industry standards (ISO/IEC, 2021) to cite a few.

Fairness in AI (or simply, *fair-AI*) aims at designing methods for detecting, mitigating, and controlling biases in AI-supported decision making (Schwartz et al., 2022; Ntoutsi et al., 2020), especially when such biases lead to (in an ethical sense) unfair or (in a legal sense) discriminatory decisions. Fairness research in human decision-making was triggered by the US Civil Rights Act of 1964 (Hutchinson & Mitchell, 2019), while bias in procedural (i.e., hand-written by humans) algorithms has been considered since the mid 1990's (Friedman & Nissenbaum, 1996)—with the first known case tracing back to 1986 (Lowry & Macpherson, 1986). Instead, fair-AI research is only 15 years old, starting with the pioneering works of Pedreschi et al. (2008) and Kamiran and Calders (2009). The area originally addressed discrimination and unfairness in ML, and it has been rapidly expanding to all sub-fields of AI and to any possible harm to individuals and collectivities. The state-of-the-art has been mainly developing on the technical side, often reducing the problem to a numeric optimization of some fairness metric (Ruggieri et al., 2023; Carey & Wu, 2023; Weinberg, 2022). Such critiques to the hegemonic (i.e., dominant) theory of fair-AI are not new to the AI community. For instance, Wagstaff (2012) questioned the hyper-focus of ML on abstract metrics "in that they explicitly ignore or remove problem-specific details, usually so that numbers can be compared across domains" but the true significance and impact of the metrics are neglected. Likewise, Mittelstadt et al. (2023) pointed out how "the majority of measures and methods to mitigate bias and improve fairness in algorithmic systems have been built in isolation from policy and civil societal contexts and lack serious engagement with philosophical, political, legal, and economic theories of equality and distributive justice", and proposed to address future discussion more towards substantive equality of opportunities and away from strict egalitarianism by default. The issue of engineering fairness is, without doubts, challenging (Scantamburlo, 2021), and likely to require domain-specific approaches (Lee & Floridi, 2021; Chen et al., 2023b) and the ability to distinguish whether and when to use AI (Lin et al., 2020), or how to enhance and extend human capabilities with AI (*human-centered AI*) (Xu, 2019; Garibay et al., 2023). A paradigmatic case is presented in Silberzahn and Uhlmann (2015), where 29 teams of researchers approached the same research question (about football players' skin colour and red cards) on the same dataset with a wide array of analytical techniques, and obtaining highly varied results. The authors concluded that "bringing together many teams of skilled researchers can balance discussions, validate scientific findings and better inform policy-makers".

The NoBIAS project

The NoBIAS project (January 2020–June 2024) was a Marie Skłodowska-Curie Innovative Training Network funded by the European Union's Horizon 2020 research and innovation program. The core objective of NoBIAS was to research and develop novel interdisciplinary methods for AI-based decision making without bias. Fig. 1 shows the NoBIAS architecture, which is designed to integrate bias management with the AI-system pipeline layer. The Bias Management Layer is made up of the various components contributed by the research projects of fifteen Early-Stage Researchers (ESRs). Together, these components aim to achieve three

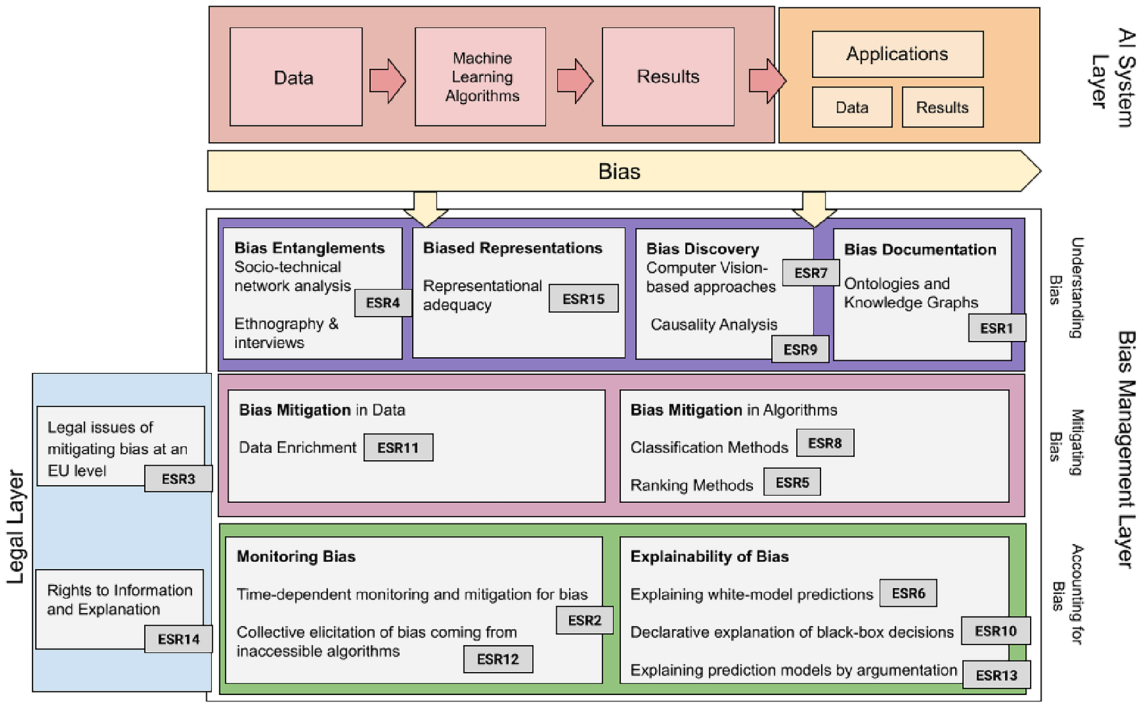

Fig. 1 The NoBIAS architecture integrates the components necessary to understand, mitigate, and account for bias, addressing the whole AI-System decision-making pipeline

main research objectives: understanding bias, mitigating bias, and accounting for bias in data and AI-systems. An orthogonal Legal Layer provides the necessary EU legal grounds supporting the research objectives. The purpose is not to produce one single bias management framework but rather to combine technologies and techniques for generating bias-aware AI-systems in different application domains and contexts.

Summary of contributions

The contributions of this paper are twofold:

- we concisely survey the state-of-the-art of fair-AI methods and resources, and the main topics about policies on bias in AI (Sect. "The landscape of policies on bias and fairness in AI"), thus providing guidance for both researchers and practitioners;
- we discuss the main policy suggestions and the best practices that, in light of the execution of the NoBIAS project, are deemed relevant and under-developed (Sect. "Lessons from the NoBIAS project"). These topics are presented w.r.t. the pillars of the NoBIAS architecture (legal challenges, bias understanding, bias mitigation, and accounting for bias).

We take a multidisciplinary approach, thus facilitating cross-fertilization.

The landscape of policies on bias and fairness in AI

In this section, we provide a concise overview of state-of-the-art fair-AI methods and policy topics. We point to the main contributions and resources in the area to provide guidance for both researchers and practitioners.

Fair-AI methods and resources

Multiple measures of the degree of (un)fairness in (automated) decision making have been introduced in ML and AI (Castelnovo et al., 2022; Mehrabi et al., 2021; Berk et al., 2021; Verma & Rubin, 2018; Zliobaite, 2017; Caton & Haas, 2024). Some of them were originally proposed and investigated in other disciplines, such as philosophy, economics, and social science (Lee et al., 2021b; Hutchinson & Mitchell, 2019; Binns, 2018a; Romei & Ruggieri, 2014). *Group fairness* metrics aim at measuring the statistical difference in distributions of decisions across social groups. *Individual fairness* metrics bind the

distance in the decision space to the distance in the feature space describing people's characteristics. *Causal fairness* metrics exploit knowledge beyond observational data to infer causal relations between membership to a protected group and decisions, and to estimate interventional consequences. As with other performance objectives, the choice of a fairness metric is crucial for optimizing AI models. See the previous surveys and Räz (2021); Wachter et al. (2021a); Hertweck et al. (2021); Binns (2020); Tang et al. (2023); Binns et al. (2023) for a discussion of the moral/legal bases and relative merits of the various fairness notions and metrics.

Fairness metrics are the building block for numerous methods and tools of fair-AI. They aim at bias detection (a.k.a. *discrimination discovery* or *fairness testing*) (Chen et al., 2022), at data de-biasing through data processing (*pre-processing approaches*) (Shahbazi et al., 2023; Zhang et al., 2023), at fair learning of AI models and representations (*in-processing approaches*) (Wan et al., 2023), at correcting existing models (*post-processing approaches*), and at monitoring models' decisions (*monitoring*) (Kenthapadi et al., 2022; Barrainkua et al., 2022). We also refer to Pessach and Shmueli (2022); Hort et al. (2022); Mehrabi et al. (2021); Ashurst and Weller (2023) and to Fabris et al. (2022); Quy et al. (2022), respectively, for surveys of the techniques and of the experimental datasets commonly used in the field. Several off-the-shelf software libraries are available to practitioners, expanding at a fast pace. Some critical gaps to be addressed by such systems are discussed in Richardson and Gilbert (2021); Lee and Singh (2021); Balayn et al. (2023). A few papers critically discuss the inherent limitations of fair-AI (Friedler et al., 2021; Buyl & Bie, 2024; Ruggieri et al., 2023; Castelnovo et al., 2023).

Research in fair-AI originated from the supervised ML area, but it has been rapidly expanding to all sub-fields of AI, including unsupervised (Chhabra et al., 2021; Dong et al., 2023) and reinforcement learning (Gajane et al., 2022), natural language processing (NLP) (Blodgett et al., 2020; Czarnowska et al., 2021; Gallegos et al., 2023), computer vision (Fabbrizzi et al., 2022), speech processing, recommender systems (Chen et al., 2023a), and knowledge representation (Kraft & Usbeck, 2022) among others. Major AI scientific conferences regularly include papers and workshops on bias and fairness. A few global events are targeted at multidisciplinary aspects of bias, fairness and other ethical issues in AI and algorithmic decision making. These include ACM FAccT[3], AAAI/ACM AIES[4], ACM EAAMO[5], and FoRC[6].

A number of initiatives have started to standardize, audit, and certify algorithmic bias and fairness (Szczekocka et al., 2022), such as the IEEE P7003™ Standard on Algorithmic Bias Considerations[7], the IEEE Ethics Certification Program for Autonomous and Intelligent Systems[8], the ISO/IEC TR 24027:2021—Bias in AI systems and AI aided decision making[9], and the NIST AI Risk Management Framework[10]. Challenges of certification schemes are discussed in Anisetti et al. (2023). Moreover, very few works attempt at investigating the practical applicability of fairness in AI (Madaio et al., 2022; Makhlouf et al., 2021b; Beutel et al., 2019), whilst several external audits of AI-based systems have been conducted (Koshiyama et al., 2021), sometimes with extensive media coverage (Camilleri et al., 2023). Finally, on the educational side, bias and fairness have become common topics of university courses on technology ethics (Fiesler et al., 2020), albeit they are not sufficiently included in core technical courses (Saltz et al., 2019) nor sufficiently transversal and interdisciplinary (Raji et al., 2021b; Memarian & Doleck, 2023).

the NoBIAS bias and fairness in AI

Bias and fairness can imply different meanings to different stakeholders depending on the application context, the people's culture and moral values, and the reference discipline (Mitchell et al., 2021; Mulligan et al., 2019). Policy initiatives, standards, and best practices in fair-AI set principles, procedures, and knowledge bases to guide and operationalize the detection, mitigation, and control of bias in AI models. Paradoxically, the uncoordinated selection and usage of fair-AI techniques may worsen off some protected groups as side-effects. Examples of such behaviors are described in the literature, including the Yule's effect (Ruggieri et al., 2023) and long-run effects of imposing fairness constraints (Liu et al., 2018).

Policy and guideline inventories

The AI Ethics Guidelines Global Inventory[11] by AlgorithmWatch lists 167 frameworks "that seek to set out principles of how systems for automated decision-making can be developed and implemented ethically". There are 8 binding

[3] https://facctconference.org/.

[4] https://www.aies-conference.com/.

[5] https://eaamo.org/.

[6] https://responsiblecomputing.org/.

[7] https://standards.ieee.org/project/7003.html.

[8] https://standards.ieee.org/industry-connections/ecpais.html

[9] https://www.iso.org/standard/77607.html.

[10] https://www.nist.gov/itl/ai-risk-management-framework.

[11] https://inventory.algorithmwatch.org/.

agreements, 44 voluntary commitments, and 115 recommendations. The EU Agency for Fundamental Rights[12] has collected a list of 349 policy initiatives at the national level, and also including examples at the EU and international level. The OECD.AI Policy Observatory[13] provides a live repository of over 800 AI policy initiatives.

An early survey of 84 ethics guidelines (mostly from Western countries) found an apparent agreement that AI should be ethical, and it identified shared principles of transparency, justice and fairness, non-maleficence, responsibility and privacy. Authors highlight, however, a "substantive divergence in relation to how these principles are interpreted, [...] and how they should be implemented" (Jobin et al., 2019). Despite these various contributions, universal standards or blueprints of fair-AI have not yet been provided by policy-makers, regulators or scientific experts (Wachter et al., 2021b). Even if there were such standards or blueprints, computer/data scientists and practitioners still need to translate these into their academic and industrial contexts and specific situations (Hillman, 2011; Kiviat, 2019).

The option not to use AI

Some scholars argue that, while AI is biased, it is less biased than humans (Lin et al., 2020). For example, humans tend to resort to judgement heuristics when making decisions, leading to biased outcomes (Kahneman, 2011). Humans can also be inconsistent and sometimes opaque and unreliable decision-makers (Kahneman et al., 2021). Given that as the alternative, the option of a noise-free, consistent algorithm is understandably appealing to some. This rationale has supported the push for algorithmic-decision-making system across domains (Miller, 2018). Notwithstanding, it is essential to acknowledge the false sense of objectivity attributed to AI as well as to revise the narrative that AI's deployment and use is inevitable. Technology alone cannot solve complex real world problems (D'Ignazio & Klein, 2020; Costanza-Chock, 2020), let alone in an equitable way (Costanza-Chock, 2020; Alkhatib, 2021). In underpinning the non-use of AI, or by-effect, prohibiting it or supporting its dismantlement, the following arguments have been documented and researched: potential or realized health and safety harms, human rights violations, opposition to deceptive predictive tools, e.g., predictive optimization (Wang et al., 2023), and organizational factors (Alkhatib, 2021). Existing community-led efforts, such as Stop LAPD Spying Coalition[14], invest their efforts in awareness campaigns on the risks and implications of the hyper surveillance of marginalized and racialized communities, thus opposing the deployment of predictive policing tools across cities. Moreover, emerging research (Pruss, 2023) has been able to demonstrate that despite the best efforts to automate high-stake decision-making, humans operating these systems can still "opt-out", or choose to not use/interact with these tools.

An underdeveloped research line consists of rejecting the low-confidence outputs of an AI system in favor of escalating the decisions to a human agent who could possibly take into account additional (qualitative) information. This is considered in the area of classification with a *reject option* (or *selective classification*) (Hendrickx et al., 2021). There is a trade-off here between the performance of an AI system on the accepted region, which should be maximized, and the probability of rejecting, which should be minimized, as human agents' effort is limited.

Regarding legal regimes, the EU law of General Data Protection Regulation (GDPR) (European Parliament and Council of the European Union, 2016) establishes some restrictions on the use of automated decision-making over individuals when the legal rights or legal status of an individual are impacted. Concretely, individuals should not be subject to a decision that is based solely on automated processing when it is legally binding or has similarly significant effects on them. Whether Article 22 of GDPR provides the data subjects with a right to object or establishes a general prohibition on automated decision-making is still uncertain and is the object of academics and practitioners debate (Mendoza and Bygrave, 2017; Article 29 Data Protection Working Party, 2018). The position of the regulator, then, seems to either offer people the option to opt-out or to provide them with strong safeguards to protect them from potential risks and harms. The upcoming EU AI Act (European Commission, 2021), will introduce in the EU legal framework a substantial advance in this regard by adopting a risk-based approach to assess AI systems' legal compliance. AI systems could only be placed in the EU market if they comply with certain requirements that mainly aim to avoid the bias. The proposed risk-based approach differentiates between minimal risk, low risk, high-risk, and unacceptable risk, advocating, likewise, for a gradually stricter set of obligations and duties proportionate to each level of risk. The AI Act bans six practices due to their particularly harmful and abusive nature that contradicts the values of respect for human dignity, freedom, equality, and the rule of law. Specifically, the text recognises the threat that AI practices concerning: (1) biometric categorization systems

[12] https://fra.europa.eu/en/project/2018/artificial-intelligence-big-data-and-fundamental-rights/ai-policy-initiatives.

[13] https://oecd.ai/en/dashboards/overview.

[14] https://stoplapdspying.org/wp-content/uploads/2018/05/Before-the-Bullet-Hits-the-Body-May-8-2018.pdf.

using sensitive attributes, (2) facial recognition, (3) emotion recognition, (4) social scoring, (5) human manipulation, and (6) the exploitation of people's vulnerabilities can pose to peoples' rights and democracy values. Notwithstanding the prohibition, the use of real-time and post-remote biometric identification systems in public accessible spaces for law enforcement purposes would be permitted under specific safeguards and strict conditions.

Using fair-AI with a guidance

Fairness metrics are at the core of the technical approaches for fair-AI. However, theoretical results state that it is impossible to satisfy different fairness notions at the same time (Chouldechova, 2017; Kleinberg et al., 2017). Not only fairness notions are in tension among each other (Alves et al., 2023), but also with other quality requirements of AI systems, such as predictive accuracy (Menon & Williamson, 2018), calibration (Pleiss et al., 2017), impact (Jorgensen et al., 2023), and privacy (Cummings et al., 2019), for which Pareto optimality should be considered (Wei & Niethammer, 2022). Moreover, the choice of a fairness metric requires to take into account several contrasting objectives: stakeholders' utility, human value alignment (Friedler et al., 2021), people's actual perception of fairness (Saha et al., 2020; Srivastava et al., 2019), and legal and normative constraints (Xenidis, 2020; Kroll et al., 2017). Decision diagrams or rules-of-thumb for guiding practitioners in the choice of the fairness metrics are offered by (Makhlouf et al., 2021a; Buijsman, 2023; Majumder et al., 2023), highlighting the complexity of the choice. The way that the various objectives and requirements are looked for, expressed and formalized, impacts on the choice of the fairness metrics and, a fortiori, on the design of an AI system (Passi & Barocas, 2019)—an instance of the *framing effect* bias, as shown e.g., in Hsee and Li (2022). For example, in the famous case analysed by *ProPublica*[15], the COMPAS algorithm for recidivism prediction fails to meet equal false positive rate among groups, but it achieves equal calibration (Corbett-Davies et al., 2017), possibly as the result of different perspectives taken by the designers of the algorithm and the ProPublica journalists. Even when restricting to a specific fairness notion, there is a problem on how to quantify the degree of unfairness. In fact, even the apparently innocuous choice among algebraic operators (e.g., difference or ratio of proportions), may have an impact. Pedreschi et al. (2012) show that the top-k protected-by-law sub-groups with the highest risk difference and the top-k with the highest selective risk ratio do not coincide. Hence, cases of possible discrimination with one choice may be undetected or unprevented with another choice.

Beyond debiasing: addressing the origins of AI harms

The report by Balayn and Gürses (2021) studies several EU policy documents, including the AI Act. The authors find that such documents rely on technocentric approaches to address AI harms, while simultaneously not adequately specifying which harms are being referred to. They argue that there is an overemphasis and overreliance specifically on the approach of debiasing data and models. Here, debiasing is used as it is in the fair-AI literature, to refer to improving model performance on specific fairness metrics, as well as to improving representation of certain groups in datasets. This is described as a limited approach as it fails to acknowledge potential harms caused by a myriad of other system design decisions, such as what is being optimized for or what attributes are being used to represent aspects of the real world. Authors also point out that the documents provide no guidance on how to address the inevitable question of which stakeholders view of what is acceptable or unacceptable bias in a system, nor do they acknowledge that any dataset or system is biased, in the sense that it was created by people, with and for a specific view or goal. They advocate for the EU to utilize other governance strategies beyond technical debiasing solutions, so as not to transfer the responsibility, and power, to determine complex political questions to designers and technicians building AI systems. One alternative perspective about the impact of AI systems they address is the organizational view. Specifically, they identify the need to consider what impacts the adoption of extensive AI systems will have on public institutions; if they begin to rely on digitization and automation helmed by large private companies, in what ways will their resources and capacities be shifted, and what would this kind of interdependence mean for public–private relationships.

Bias and auditing

In algorithmic decision-making, auditing involves using experimental approaches to investigate potential discrimination by controlling factors that may influence decision outcomes (Romei & Ruggieri, 2014). Given the application scope of these systems, proposed audits span various domains, including algorithmic recruitment (Kazim et al., 2021), online housing markets (Asplund et al., 2020), resource allocation systems (Coston et al., 2021); and more general processes related to the design (Katell et al., 2019) and vision of these systems as socio-technical processes (Cobbe et al., 2021). Auditors play a crucial role in ensuring algorithmic accountability. Consequently, they involve multiple stakeholders, from product developers, government,

[15] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

policy makers, and data owners to broader groups in society, such as advocacy organizations and institutional operators (Wieringa, 2020). Ultimately, audits are evaluations designed to hold stakeholders accountable. Algorithm auditing (Koshiyama et al., 2021), and specifically AI auditing (Mökander, 2023), is a concept coined to seek for the development of auditing frameworks on research and *in practice*. Moving from a case-by-case basis, audits should establish formal assurance that algorithms are legal, ethical, and safe by informing on governance and compliance with regulations and standards. Notably, the Information Commissioner's Office (ICO) of the United Kingdom has developed a such a framework for auditing AI systems in the public and private sectors[16]. These investigations assess how these entities process personal information and effectively deal with information rights issues. In this capacity, an audit will involve a thorough evaluation of an organisation's procedures, processes, records, and activities. We see in this example how audits are crucial in addressing issues of bias and discrimination. Specifically, by ensuring the existence of adequate policies and procedures, verifying their compliance, testing the adequacy of controls, detecting existing or potential violations, and recommending necessary changes to controls, policies and procedures.

Living with bias by documenting it

An emerging scholarship advocates for the development of documentation practices and accompanying artefacts that enhance AI audit pipelines (Gebru et al., 2021; Raji & Yang, 2019; Stoyanovich et al., 2022; Raji et al., 2020), thus enabling stakeholders to easily inspect all the actions performed across the many steps of the pipeline. This also contributes to increasing the trust on the development processes and the systems themselves. The AI community does not count with standardized methods to produce documentation on datasets and models, nor are there any specific regulatory frameworks that enforce this practice at the moment of writing; however, pioneering work in this area argues that "drawing on values-sensitive practices can only bring about improvements in engineering and scientific outcomes" (Bender & Friedman, 2018). Further, Gebru et al. (2021) advocate that documentation promotes the communication between "dataset consumers and producers". Existing frameworks for the elaboration of documentation include: Datasheets for Datasets (Gebru et al., 2021), Dataset Nutrition Labels (Chmielinski et al., 2022), Data Statements (Bender & Friedman, 2018), Data Readiness Report (Afzal et al., 2021), and Model Cards for Models (Mitchell et al., 2019). Formal data models, like ontologies and controlled vocabularies, can also support AI-related documentation needs. Examples of relevant vocabularies include: the Data Catalog Vocabulary (DCAT[17]), the provenance ontology (PROV-O[18]), and the Machine Learning Schema ontology (MLS[19]). Lastly, Miceli et al. (2022b) propose a shift in perspective, from documenting datasets to documenting data production processes in order to account for the intensive and precarious human labour involved in the production of datasets. More recently, the urgent call for data stewardship (Peng et al., 2021) and responsible data management practices (Stoyanovich et al., 2022) has also seen the emergence of new professional roles (Rismani & Moon, 2023).

Lessons from the NoBIAS project

The Bias Management Layer in the NoBIAS architecture of Fig. 1 aims at achieving three main research objectives: understanding bias, mitigating bias, and accounting for bias in AI-based systems. An orthogonal Legal Layer provides the necessary legal grounds, with regard to the EU context, supporting the research objectives. In this section, we discuss a few policy advices and best practices resulting from the execution of the NoBIAS research. The section is organized according to the NoBIAS architecture.

Legal challenges of bias in AI

After framing the EU legal context of AI biases, we discuss how to overcome the hegemonic theory of fair-AI beyond fairness metrics by moving towards transparency and accountability of AI systems. Finally, we consider the synergies and frictions between non-discrimination and data protection law in the specific case of EU legislation. A summary of the challenges, policy advices, and best practices in the Legal Layer is reported in Fig. 2, together with a reference to the subsection(s) where they are discussed.

AI biases, discrimination and unfairness

Anti-discrimination legal cases—targeted and strategically litigated—are traditionally based on causal connections between the protected group, the questioned provisions, and the discriminatory situation or unfair treatment (Foster, 2004). However, AI systems challenge that, initial, intuitive causality

[16] https://ico.org.uk/media/for-organisations/documents/4022651/a-guide-to-ai-audits.pdf.

[17] https://www.w3.org/TR/vocab-dcat-2/.

[18] https://www.w3.org/TR/prov-primer/.

[19] https://github.com/ML-Schema/core.

Legal Layer

- AI models often lack the auxiliary causal knowledge required to prove anti-discrimination cases as these require to show that decision is *because of* (i.e., *at cause of) the protected ground.* (AI biases, discrimination and unfairness)
- AI models' complexity and opaqueness make it difficult to identify individuals and groups that are treated unfairly. (AI biases, discrimination and unfairness)
- The design of AI models requires to agree on and to operationalize legal and ethical principles. (AI biases, discrimination and unfairness)
- Transparency and accountability of AI systems are a way to overcome the hegemonic theory of fairness, which reduces the fairness problem to quantitative metric optimization. (AI fairness beyond metrics: transparency and accountability of AI systems)
- There are synergies and frictions in the EU legal framework between data protection law and non-discrimination law, which demand for an integrated and interdisciplinary techno-legal framework of bias management. (EU data protection law and non-discrimination law)

Fig. 2 Legal Layer: challenges, policy advices, and best practices. In parenthesis, references to the subsections where they are discussed

basis by performing through correlations that do not provide causal explanations for the connections between the input data and the target variable (see Bathaee (2018) and also later Sect. "Bias as a causal-thing"). AI systems operate in such a complex manner that they defy human understanding, leaving the potential victim unaware of the scope and magnitude of the extent to which they have been discriminated against and disadvantaged. Establishing a case of AI discrimination is undoubtedly difficult, as seen in the following brief analysis. Firstly, the potential claimants may not be aware of their disadvantage and the information required to prove that such algorithmic discrimination may be difficult to discover, gather, or access (Wachter et al., 2021b). Secondly, anti-discrimination law protects on the grounds of protected attributes; however, the sources of algorithmic discrimination and the individuals and groups affected by it may not be straightforwardly correlated with those attributes (Zuiderveen Borgesius, 2020). Protected groups may be treated in a biased or unfair way, but the use of proxies can cover such treatment as the features of the model would not directly reveal the use of any sensitive attribute. A second challenge arises from the limited personal scope of EU non-discrimination law, restricted by an exhaustive list of protected grounds. By utilizing proxies-i.e., "neutral" variables closely correlated with the protected ones-the use of AI systems poses a significant risk of circumventing the scope of legal protection [often referred as proxy discrimination (European Commission et al., 2021; Zuiderveen Borgesius, 2020)]. The way AI systems operate reinforce an existing challenge in EU equality protection, that of intersectional discrimination, arising when discriminatory effects occur at the intersection of two or more vectors of disadvantage. While concepts of intersectionality have been advanced by legal scholarship, the Court of Justice of the EU has so far failed to explicitly recognize intersectional discrimination as a special type of discrimination (Xenidis, 2018; Roy et al., 2023; European Court of Justice, 2016), creating a potential gateway for algorithmic discrimination within the realm of EU non-discrimination law. Thirdly, the current legal procedure to establish a case of discrimination may also set some limitations to bring and present a case of algorithmic discrimination effectively (Wachter et al., 2021b). Furthermore, what makes an algorithm biased and its outcomes unfair is the subject of a contested debate (Rovatsos et al., 2019; Barocas & Selbst, 2016; Jacobs, 2021; Wachter et al., 2021a). Fairness is essentially a contested concept as it is context-dependent and highly conflicts with different ethical, political, and cultural understandings. Still, fairness needs to be mathematically defined to build fair-AI systems, leaving the question of which values need to be operationalized into variables unsolved. For this reason, the literature of fair-AI mainly derives its fairness constructs from a legal context where a process or decision is considered fair if it does not discriminate against people based on their membership to a protected group (Tolan, 2019; Mehrabi et al., 2021; Romei & Ruggieri, 2014). Fairness can be understood as equality or as equity, which are different concepts (Minow, 2021), so the instruments and ways to achieve and ensure the goals of each highly differ. Fairness, in essence, can be understood in different manners depending on its nature, formal or substantive; the context it applies to, legal or technical, or the actor it refers to, public or private. Selecting the appropriate principles and operationalizing the preferred construct requires understanding how people assess fairness and questioning whose perceptions should be captured or discharged (Binns, 2018b).

AI fairness beyond metrics: transparency and accountability of AI systems

Carey and Wu (2023); Weinberg (2022) survey the existing critiques on the hegemonic theory of fairness that draw from non-computing disciplines, including philosophy,

law, critical race and ethnic studies, and feminist studies. The hegemonic (i.e., dominant) theory of fairness in the ML community reduces the fairness problem "in terms of a domain-general procedural or statistical guideline [...] so long as the chosen fairness criteria are satisfied, the resulting procedures and outcomes of the system are necessarily fair" (Green & Hu, 2018). Beyond those critics, AI systems' opaqueness and the potential to impact individuals' lives are frequently described as the main motivations to demand disclosures of information and provision of explanations about their internal processes and final outcomes, understanding these requirements as necessary to ensure effective governance of the AI context (Almada, 2021) and for allowing applicants to make cases of discrimination (Xenidis and Senden, 2020). On the one hand, algorithms are considered powerful procedures that create "a growing need to evaluate the claims, decisions, actions, and policies that are being made on the bases of them. This evaluation requires gauging the reasons for an algorithmic decision, its components, and the weight assigned to them" (Vedder & Naudts, 2017), in short, requiring *AI accountability*. On the other hand, the "individual adversely affected by a predictive process has the right to understand why and frames this in familiar terms of autonomy and respect as a human being" (Edwards & Veale, 2017), in short *AI transparency*.

An extensive review of algorithmic accountability is provided by Wieringa (2020), while Percy et al. (2021) brings to life the notion of AI accountability in industry work programs, aiming to implement industry-specific technical requirements. Algorithmic impact assessments are accountability governance practices rendering visible the (possible) harms caused by algorithmic systems (Metcalf et al., 2021). Reviewability, introduced by Cobbe et al. (2021), is a way to break down the algorithmic decision-making process into technical and organisational elements which help in determining the contextually appropriate record-keeping mechanisms to facilitate meaningful review both of individual decisions and of the process as a whole. The design of interpretable AI models and the development of methods to explain black box models are comprised in the area of *eXplainable AI* (XAI) (Guidotti et al., 2019; Minh et al., 2022). Such techniques respond to a societal desire to understand the obscure systems that can greatly affect our lives when allocating services or granting and denying rights. Transparency and information obligations can publicly assess the consistent compromise and dutifulness of AI systems with legal principles such as fairness, lawfulness, or information privacy, improving the legitimacy and acceptance of their use by the individuals affected by them at last stay, and supporting the contestability of their outcomes (Henin & Métayer, 2022). However, in most situations where there are obligations to provide information and explanations about automated decision-making systems, the context is adversarial, and the interests of the parties involved are, if not opposite, different (Bordt et al., 2022). The interest of the users and providers of AI systems and the persons affected by them are opposed to the extent that the former will want to address its transparency and information obligations in a way that ensures compliance but does not harm its private interests, whilst the person subjected to the AI systems will expect a level of compliance that is sufficiently rigorous to enable an effective exercise of her rights and protect her interests and freedoms. Consequently, the interests to be protected or respected will largely condition the method of explanation and the information and explanations expected to be received (see also later Sect. "The need for trustworthy AI, and XAI in particular").

EU data protection law and non-discrimination law

The uptake of (fair-)AI has brought two distinct EU legal regimes to the forefront: data protection law and non-discrimination law. As data-driven technology, AI relies today on the processing of big volumes of data, which often relate to identified or identifiable individuals. This processing brings the development and deployment of many AI systems directly under the scope of the GDPR (European Parliament and Council of the European Union, 2016). On the other hand, due to the issue of bias, AI applications have the potential to infringe upon non-discrimination rights and interfere with existing non-discrimination regulations. Considering that both data protection and non-discrimination rights constitute fundamental rights that are as such equally protected in EU primary (art. 8 and 21 of the Charter of Fundamental Rights (European Union, 2000)) and secondary law (GDPR and EU non-discrimination directives (Council of the European Union, 2000a; European Parliament and Council of the European Union, 2006; Council of the European Union, 2000b, 2004)), mapping aspects of their intersection becomes highly relevant. We refer to Gellert et al. (2013) for a comparative analysis of the two. Here, we highlight a few relevant synergies and frictions.

Since the emergence of the AI bias discourse, EU legal scholars have approached the existing non-discrimination and data protection legal frameworks in an integrated way in order to deal with the challenges of AI in the digital age (Zuiderveen Borgesius, 2020; Hacker, 2018; European Parliament et al., 2022). Confronted with the novel challenges of algorithmic bias, commentators have mainly sought recourse to the GDPR, as a means to compensate enforcement deficiencies of the EU non-discrimination legal apparatus. Tools such as individual access rights (Article 15 (1)), data protection audits (Article 58 (1) (b)), Data Protection Impact Assessments (Article 35 et seq.) and the principle of "fairness" (art. 5 (1) (a)) along with the provision of administrative fines for violation of associated obligations (art. 83) are

among those highlighted for their potential to fight against AI bias and support the protection of non-discrimination rights. However, recourse to data protection law cannot be forever a panacea for the challenges of AI discrimination. Not only is the GDPR not rationae materiae *primarily* concerned with the right to non-discrimination but it is also de facto considerably ineffective in achieving this goal (Zuiderveen Borgesius, 2020; European Parliament et al., 2022). It is important that EU and national legislature and judiciary engage with the limitations of existing non-discrimination frameworks and the nuances of AI application in order to consider tailored legislative amendments or interpretative approaches. Specific recommendations or guidelines by relevant independent bodies such as the European Data Protection Board (EDPB) that adapt the application of existing legislation to the specificities of AI technologies will particularly serve this effort. Striking the right balance between legal certainty and agile application across different domains, Member States and technological developments represent a key challenge in this undertaking. See Gerards and Zuiderveen Borgesius (2022); European Parliament et al. (2022); Xenidis (2020) for suggestions on different legislative and interpretative approaches in the context of fair-AI.

The fair-AI ecosystem may bring about a clash between the objectives of data protection and non-discrimination legislation, as debiasing approaches may interfere with well-established data protection rights and principles (Veale & Binns, 2017). First of all, the lack of representative training datasets has been consistently described as one of the sources of AI bias (Barocas & Selbst, 2016; Buolamwini & Gebru, 2018; Ntoutsi et al., 2020) (see also Sect. "Understanding bias"). This line of reasoning has been adopted by the proposed AI Act (European Commission, 2021). Specifically, art. 10 para 3 mandates that providers of high-risk AI systems shall ensure representative training, validation and testing data sets, as part of the prescribed data governance practices. It is thus conceivable that such legislative calls might risk motivating an increasing collection of personal data particularly from data subjects that belong to hitherto underrepresented groups, who are often the most vulnerable in terms of data protection. Furthermore, fair-AI frameworks centered around bias detection, monitoring, and correction often imply the processing of data on characteristics protected by the EU non-discrimination law. This often corresponds to the collection and/or the processing of special categories of personal data (hereafter sensitive data), despite the fact that they are, as such, extensively protected by the GDPR (European Parliament and Council of the European Union, 2016). Moreover, special attention must be given to the way that bias mitigation approaches, and particularly the modification of training data through pre-processing (see Sect. "Fair-AI methods and resources"), may interfere, or at least may introduce a layer of complexity, with GDPR principles such as the principle of "accuracy" outlined in Article 5(1) (d) of GDPR.

Since the practice of removing or ignoring sensitive attributes shows to be ineffective to tackle the issue of AI bias (Barocas et al., 2019; Zliobaite and Custers, 2016; Haeri & Zweig, 2020), data scarcity due to regulation constraints is essentially seen as a hurdle to the realisation of fair-AI. There is an effort in the European Parliament's negotiated version of the AI Act to minimize and circumscribe the width of this obligation, by requiring *"sufficiently* representative" (*sic*) training datasets. However, this choice can also be seen critically as compromising and relativizing the obligation of AI providers to engage with representation biases. As the notion of "sufficiency" is not legally defined and until specific standards or guidelines elaborate on the matter, it is at the discretion of AI providers to weight up their datasets against the "sufficiency" scale, considering the application and the context at hand. A level of legal uncertainty arises in that regard.

The proposed AI Act comes to mediate this tension and opens up the possibility of processing sensitive personal data for the case of bias monitoring, detection and correction in high-risk AI systems [art. 10 (5)]. This possibility comes together with various requirements, intended to ensure a balance between the right to data protection and non-discrimination and prevent an excessive processing of sensitive data in the name of debiasing. However, once again these requirements entail indefinite legal concepts (e.g. "necessity"), with no existing guidance on they way they shall be operationalized in the context of fair-AI. Entrusting the lawful interpretation and implementation of fundamental requirements to the discretion of AI providers entail the risk of a purposeful and inconsistent legal application to the detriment of the right to data protection. In addition, infringements upon provisions of the GDPR or the AI Act might result in severe financial penalties (art. 83, 84 GDPR, art. 71 AI Act).

The tensions between different regulatory tools and the abundance of vague binding textual requirements generate thus a great level of legal uncertainty for all bodies concerned, which explains the urgent need for adequate guidance. Considering the novelty, the fast-evolving nature and the complexity of different debiasing approaches, the desired guidance requires targeted research efforts. Rather than focusing solely on non-discrimination desiderata and sustaining an adversarial conceptualisation of "fairnes" vs "privacy", it is important that interdisciplinary research and good practices on fair-AI transition to a more integrated model. This model should account for the deep intertwinement between data protection and non-discrimination legal regimes and seek to enhance privacy while engaging in debiasing.

Bias Management Layer - Understanding Bias

- We should acknowledge that there are many forms of bias, with different roots and effects. (Understanding biases, not bias)
- The "ground truth" is a myth. It does not exist in a structurally unjust and unequal society. (The ground truth is biased)
- Data curation in AI should import source criticism and archival practices from historical and humanistic disciplines. (Beyond documenting bias: source criticism and archival practices)
- There is an hyper-fixation on data as the primary source of bias, but the whole AI pipeline needs to be addressed, including the data annotation process and data labourers' exploitation. (Don't blame the data, don't blame the annotator)
- Different data types require specific regulatory guidelines and standards. (Consider the data type)

Fig. 3 Bias management layer—understanding bias: challenges, policy advices, and best practices. In parenthesis, references to the subsections where they are discussed

Understanding bias

Bias in data is not as clear-cut as it is often presented. What we mean by bias, what we consider its sources, and what we view as its materialization are all, among other, complex questions with considerable implications on policies for addressing unfair AI models. In this section, we present different angles to better understand and be critical about bias(es) in data. First, we argue on understanding biases, not bias, as a multifaceted issue. Then, we criticize the AI assumption of ground truth, quest for source criticism and archival practices, and discuss the issue of reliable data annotation. Finally, we claim for approaches specific to data types and domain types. A summary of the challenges, policy advices, and best practices in understanding bias is reported in Fig. 3.

Understanding biases, not bias

Bias is primarily understood as a difference between what is seen as "truth" or "fact" and the respective results of an algorithmic function (a prediction or a representation). Such definitions of bias have in common that they do not relate to the harmful and discriminative impact of statistical errors nor to the underlying social conditions leading to bias. Recent research not least in Computer Science has therefore elaborated how bias is also entangled with social and historical prejudice and discrimination. For instance, the terminology "gender bias" refers not only to a statistical error but also to the algorithmic amplification of already existing discrimination against women and LGBTIQ* persons like in the case of the Austrian public employment service algorithm (Lopez, 2019). Further studies grounded in Social Science and Science and Technology Studies have explored the "empirical grounded accounts of practices" (Jaton, 2020) of Computer Science, folded into algorithmic bias and fairness. These contributions have in common that they approach bias not as a statistical error in the predictive performance of an algorithm, but as socio-technical, procedural and constitutive to algorithms (Jaton, 2020; Ziewitz, 2016; Draude et al., 2019; Seaver, 2017).

We think it is crucial to acknowledge that there is not just a singular bias, but rather a multitude of biases, having different (social, technical and socio-technical) roots and exerting distinct effects when employed. In the realm of policy-oriented research, a suggested approach is to "study up" (Nader, 1972) and embrace a framework that considers power dynamics, rather than solely focusing on identifying (singular) bias(es) (Miceli et al., 2022a). By doing so, *understanding biases* can even inform policy-making as it acts as a synecdoche for structural inequalities that persist in society.

The ground truth is biased

AI models are trained on historical data to accomplish a certain task, e.g., to predict recidivism of defendants. The data used for training is assumed to encode the ground "truth" of the task, e.g., the actual outcome of recidivism for each defendant in case the defendant would have been released. In most cases, collecting the ground "truth" is difficult, expensive, or even unethical, as it would require to obtain counterfactual outcomes, such as releasing potential criminals, not treating sick patients, etc. (Tal, 2023). In the analysis of the COMPAS algorithm, for example, ground truth was approximated by the actual re-arrest outcome of defendants in the 2 years period after they were scored. First, due to unobservability of crime, re-arrest does not coincide with re-offense (Bao et al., 2021), which is the recidivism outcome intended to be predicted. Second, we do not know whether or not defendants who were not released would have recidivated in case they would have been released. Similarly, we do not

know whether an applicant with denied credit would have repaid the credit if granted, a sample selection bias problem tackled by reject inference in credit scoring (Ehrhardt et al., 2021). An idea close to reject inference has been considered in (Ji et al., 2020) for group fairness. Such sampling bias in collected ground truth has been called *negative legacy unfairness* (Kamishima et al., 2012), or the *selective label problem* (Lakkaraju et al., 2017), and it is an instance of data missingness (Goel et al., 2021). Recognizing that ground truth in collected data is biased help to solve the illusive tension between fairness and accuracy (Wick et al., 2019). In NLP, the ground truth is obtained by human annotation, typically aggregating annotators' labels through majority voting. Here, the simplifying assumption of a *single* ground truth is used. A perspectivist approach is emerging in favor of granting significance to divergent opinions, by designing methods over non-aggregated data (Cabitza et al., 2023). Uncertainty and inconsistency in expert annotations have been pointed out also in the domain of healthcare (Lebovitz et al., 2021; Sylolypavan et al., 2023). In the absence of unbiased ground truth, however, practitioners train AI classification models by setting the target feature using historical data. Any bias in the historical data risks to be lifted to the AI model with a false claim of fairness. Looking at other disciplines, Zajko (2022) points that AI students are untrained and unprepared for the reality of an unfair society. We support the author's claim that "AI developers refer to the reality that exists outside of their models as the 'ground truth', and bias is often defined as deviations from this truth, or inaccurate representations and predictions. But when the truth is that society is deeply, structurally unjust and unequal, and that technologies are part of these structures, the question is whether our algorithms should accurately reproduce inequality or work to change it".

Beyond documenting bias: source criticism and archival practices

Data curation is central in Computer Science approaches to data bias management (Demartini et al., 2023; Balayn et al., 2021) and information resilience (Sadiq et al., 2022). Here, we highlight instead the issue of source criticism, which is central in historical disciplines and in the humanities, but still in its infancy in Computer Science and AI. Source criticism relates to the practice of understanding the provenance, authenticity, and completeness of sources used in scholarship (Koch and Kinder-Kurlanda, 2020). In the historical disciplines and in the humanities more generally, the practice is considered as required for assessing the validity and reliability of findings based on the source, usually a document. The adoption of source critical practices, applied to datasets, in fair-AI would allow us to give a better picture of the data being used and the individual instances it contains. Questions of provenance, which is defined as "the question of who has created it with what intention, in which institutional and socio-cultural context" (Koch and Kinder-Kurlanda, 2020), have gone particularly under-examined in AI research and development work. There is now a growing body of works examining the lack of quality, offensiveness, and un-curated nature of some of the massive datasets used for common text and image AI applications (Birhane & Prabhu, 2021; Birhane et al., 2021) as well as works attempting to identify the 'genealogy' of commonly used datasets and benchmarks, with a focus on understanding the norms and values embedded in them (Raji et al., 2021a; Denton et al., 2021).

Many existing datasets used in fair-AI research have minimal information available about the reasons and decisions behind their creation (Fabris et al., 2022; Quy et al., 2022), which are needed for effective source criticism. There have been recent works proposing specific implementations for ensuring that newly created datasets are both well documented and designed as suitable for their intended purposes. In this way, AI practitioners will have a better understanding of the provenance, authenticity, and completeness of the datasets that they use, and of what the implications of results drawn from them are. Hutchinson et al. (2021) present a framework for dataset creation drawn from software development best practices. This framework is intended to support transparency and accountability regarding all steps of the dataset creation life cycle, with a particular focus on the often forgotten *maintenance* phase. Jo and Gebru (2020) propose the creation of an interdisciplinary sub-field of dataset archiving as a way to ensure capacity for the extensive and specialized work required for responsible data creation and management. The authors explain that the existing field of archiving already has established standards and practices for responsible archival processes that can be transferred to this new sub-field.

Don't blame the data, don't blame the annotator

The current paradigm of AI research and development is heavily dependent on data. Consequently, and despite the extensive resources that have been allocated to research pertaining to bias detection and mitigation in datasets and AI models, the common misconception that bias originates in the data persists, especially in circles outside fair-AI research. The hyper-fixation on data as the primary source of bias can wrongfully lead to treating the negative societal impacts of ML-systems' deployment as an oversimplified problem that can be tackled by "removing" bias from data. Instead, it is essential to reinforce the need to assess algorithmic harms through a holistic assessment that contemplates the whole of the AI pipeline throughout its entire life cycle, whilst also accounting for the societal context for its

intended use (Suresh & Guttag, 2021). With this in mind, we reinstate how biases can arise at any point of the pipeline as they are derived from the series of choices and practices that go into making these systems, and that eradicating all the biases is a near impossible task (Olteanu et al., 2019). Suresh and Guttag (2021) propose a framework that supports the understanding of sources of harm that can be mapped to different stages across the ML life cycle, accompanied by a non-comprehensive taxonomy of biases that can be attributed to each stage. Here, we emphasize on *non-comprehensive*, because in the same way humans are plagued by innumerable types of biases, datasets and models are also subjected to this problem (Olteanu et al., 2019).

Ultimately, decoupling the AI pipeline in stages can support the careful examination of harms, and help anticipate unforeseen negative implications that these technologies can go on to have upon deployment. Moreover, assessing algorithmic harms from a holistic point of view, also instils a degree of accountability from all those involved in the process of deploying them, instead of doing away with it by simply tackling bias during data pre-processing.

Another localized issue associated with the need for vast amounts of annotated data to train and validate AI-powered systems, in particular those resorting to supervised ML methods, is the one concerned with attributing data bias and, consequently, bad dataset quality, to human annotators, or by-effect, data labourers (Li et al., 2023). In particular, research focused on crowdsourcing dataset annotations tend to make the case for bias in human annotations as being one of the main causes of unfairness in downstream ML tasks (Demartini et al., 2023). The reason for this can be closely intertwined with the interpretative nature of tasks such as data labelling (see also Sect. "The ground truth is biased"), where data labourers are expected to fit complex and divergent world-views into rigid categories (Lin and Jackson, 2023). However, identifying "annotator bias" as the root problem of biased datasets, has become as of late a contentious issue in discussing ethical practices and AI development, as it overlooks the need to acknowledge opaque dataset production processes that require an intensive amount of human labour[20]. Emerging research on this spectrum calls to instead consider biases in datasets as the result of "instruction bias" (Parmar et al., 2023), where bias enters the data collection process at the hand of those designing the instructions for the requested tasks (requestors). Going further than that, Miceli et al. (2022b) propose shifting the discussion away from "annotator bias" altogether, and instead towards the critical assessment of existing work practices and conditions associated with dataset production. Specifically in this context, they allude to their restricted ability to ask questions in instructions, raise concerns about tasks, low pay, and the elevated surveillance of the labourers. To alleviate this, Miceli et al. (2022b); Li et al. (2023) advocate for centring data labourers' well-being, and propose frameworks that, for starters, incorporate their input and feedback into production processes, with the aim to empower them.

Consider the data type

We have already displayed how bias is an umbrella term that comprehends many different characterisations and ranges across different disciplines (e.g., Statistics, Psychology, Social Science, Science and Technology Studies, Gender Studies, etc.), as further demonstrated by the extensiveness of the projects[21,22] that try to catalogue human biases. Especially for big (non-tabular) data, there is a great amount of different biases that can co-occur in the same dataset and often depend on the data type itself. In visual data, for example, framing bias is defined in Fabbrizzi et al. (2022) as "any associations or disparities that can be used to convey different messages and/or that can be traced back to the way in which the visual content has been composed". It is clear how this definition makes sense only if we rely on further knowledge on how visual communication works (also from the very practical point of view). Furthermore, a typical example of bias in hate speech detection is that African American English (AAE) tends to be labelled as offensive (Harris et al., 2022). Outside the specific example of this case study based on Twitter data, for which the bias was due to a different use of swearing by AAE speakers, it is evident how searching for such a bias in general is not straightforward and requires a certain understanding of how languages work and of the relationships between different dialects of the same language. It is to be considered a best practice, then, to analyse data in search for bias having clear the peculiarities of each data types. Furthermore, any policies that aim at regulating AI adequately need to be either general enough to comprehend the specificity of each data type or differentiate among different data types. For example, the "horizontal"[23] data governance approach of the AI Act w.r.t. bias in training, testing and validation datasets (art. 10 of AI Act) might raise considerable challenges in that respect. While different types of data imply different challenges in terms of fairness and data protection, horizontal legal requirements lean arguably towards the paradigm of tabular data. This

[20] https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/.

[21] https://catalogofbias.org/biases/.

[22] https://en.wikipedia.org/wiki/List_of_cognitive_biases.

[23] "Horizontal" is to be understood here as applying uniformly to any training, testing and validation dataset used in high-risk AI systems irrespective of the data-type.

Bias Management Layer - Mitigating Bias

- Multi-stakeholders participatory design and human-centered AI can be a valid alternative to technological solutionism. (Multi-stakeholder participatory design, Prioritizing Human-centric AI)
- Intersectionality requires special attention and specific methods to account for the interplay of the different (protected) attributes. (Intersectionality)
- A principled way of tackling bias is to rely on causal reasoning. (Bias as a causal-thing)
- Relying exclusively on raw data for a given task is often not sufficient. External sources can support the so-called knowledge-intensive tasks. (Knowledge-informed AI models)
- There is an urgent need for expanding the fair-AI research on the non i.i.d. case. (The non i.i.d. case: bias in unsupervised learning and graph-mining)

Fig. 4 Bias management layer—mitigating bias: challenges, policy advices, and best practices. In parenthesis, references to the subsections where they are discussed

might impede their consistent application to a large amount of high-risk AI systems that utilize visual data. The development of corresponding regulatory guidelines and standards tailored to different data types can increase legal certainty and enhance compliance.

Mitigating bias

Bias mitigation is a crucial aspect in the development of fair-AI models, aimed at reducing or eliminating biases that can skew outcomes and perpetuate discrimination. As mentioned in Sect. "Fair-AI methods and resources", bias mitigation can happen in multiple crucial stages, including data processing approaches (*pre-processing*), specialized fair-AI algorithms (*in-processing*), and model sanitization (*post-processing*). The effectiveness of mitigations at those stages presents some challenges. Pre-processing approaches may inadvertently remove relevant or informative data, with the risk of overgeneralizing and ignoring legitimate differences that may exist among subgroups. This is a problem shared with the data processing for privacy-enforcement (Shahriar et al., 2023). In-processing approaches follow the optimization of some trade-off between performance and fairness metrics. Finally, post-processing approaches may not address the root causes of biases, hence having a limited impact and potentially leading to new biases or feedback loops. In this section, we present issues that deserve specific attention by the practitioners when implementing bias mitigation strategies. A summary of the challenges, policy advices, and best practices in the mitigating bias is reported in Fig. 4.

Multi-stakeholder participatory design

As observed in Sect. "Don't blame the data, don't blame the annotator", every technical decision, yet apparently-neutral, in any stage of the AI pipeline can impact on the biases of the final AI system. For instance, fairness is affected by imputation of missing values (Caton et al., 2022), by encodings of categorical features (Mougan et al., 2023), by feature selection strategies (Galhotra et al., 2022), and even by hyper-parameter settings (Tizpaz-Niari et al., 2022). More importantly, the composition of data transformations and AI models that are fair in isolation may not be fair in the end (Dwork & Ilvento, 2019). Observe that this also applies to AI-based complex socio-technical systems resulting from the composition of AI, algorithms, people, and procedures (Kulynych et al., 2020). The lack of compositionality requires that the bias analysis of a socio-technical system is conducted as a whole, not by pieces. This is also because the objectives and requirements of the designers of AI, of the users of AI, and of the population subject to the AI decisions are unlikely to be the same. Fair-AI methods are currently not sufficiently robust and they can be incomplete in modelling the complexity and dynamic of the deployment scenario. Multi-stakeholders participatory design (Feffer et al., 2023) and policy actions that take into account qualitative contextual information and feedback from reality may be a valid alternative to technological solutionism. For instance, the NoBIAS project contributed in Scott et al. (2022) to a participatory approach in the design of algorithmic systems in support of public employment services.

Prioritizing human-centric AI

In addition to the issues discussed in Sect. "Multi-stakeholder participatory design", involving the interested communities during the whole development process of a decision-making system is also a crucial aspect for prioritising AI systems that respond to human values—an objective known as *AI alignment* (Ji et al., 2023) or *socially responsible AI* (Cheng et al., 2021). Inclusion should go beyond the provision of "low-resource" methods (Gururangan et al., 2022), i.e., framing the under-representation of social minorities as a data scarcity problem. Instead, it should account for preventive considerations that respond to diverse human needs and preferences.

This concept is the basis for a *human-centered AI* (Mosqueira-Rey et al., 2023; Xu, 2019; Garibay et al., 2023). Active participation during the whole construction process of an AI system can be a key part of addressing the representation bias that prevails in current systems. Involving a diverse group of people has shown to be critical in stages such as selecting the preferences instructed to the model to make decisions (Organizers Of QueerinAI et al., 2023). Such practices elucidate how systems align with values from specific social groups, which frequently reflect structural and power inequalities. Adjusting to and uncovering the variations on how the data captures under-represented communities can help to represent them more fairly. For example, these practices can help to build socially aware language technologies that are adept for different dialects (Ziems et al., 2022) (see also the AAE example in Sect. "Consider the data type"). Further examples will be considered in Sect. "The need for trustworthy AI, and XAI in particular".

Intersectionality

Many bias mitigation techniques assume in input the specification of one or more protected attributes to mitigate the bias against. However, different dimensions of identity cannot be understood in isolation but must be considered collectively to grasp the full complexity of individuals. A special effort must hence be employed to consider the interplay of the different (protected) attributes (Ovalle et al., 2023). It is further worth noticing that debiasing for a group can reduce even more the representation of already under-represented subgroups (Smirnov et al., 2021). The phenomenon of debiasing paradox (Smirnov et al., 2021; Hughes, 2011), refers to situations where efforts to reduce bias towards certain groups based on a characteristic can actually exacerbate the underrepresentation of already marginalized or even the most marginalized subgroups. This paradox arises when additional attributes, which may be sensitive but overlooked or disregarded, are associated with the characteristic being targeted for bias reduction. Such correlations can occur naturally in real-life scenarios. For instance, the gender pay gap, which can be partially attributed to the wage penalty for motherhood (Budig & England, 2001), serves as an example. In this case, two attributes, namely "being a woman" and "taking care of children" are correlated and both can have detrimental effects on salary. Attempting to address bias solely based on gender may unintentionally disadvantage certain minority groups, such as women without caregiving responsibilities or men who do have such responsibilities. Hence, when considering mitigation strategies, side effects on different subgroups should be carefully analyzed. Beyond its legal challenges (see Sect. "AI biases, discrimination and unfairness"), intersectionality is currently actively addressed also by technical research (Gohar & Cheng, 2023) and Science and Technology Studies (van Nuenen et al., 2022).

Bias as a causal-thing

As observed in Sect. "AI biases, discrimination and unfairness", most ML models are purely observational and rely on correlation among features. Consequently, they are not able to account for spurious effects. A principled way of tackling bias is to rely on causal reasoning (Nogueira et al., 2022; Spirtes & Zhang, 2016). The preferred causal framework used within ML is that of Perlian Causality, or Structural Causal Models (SCM) (Pearl, 2009). Under SCM, causes and effects among a set of variables are denoted using a directed acyclical graph (DAG) that, in turn, represents a set of structural equations that encode directed effects (i.e., $X \rightarrow Y$ for attributes X and Y) rather than non-directed effects (i.e., $X \rightarrow Z \rightarrow Y$ for one or more intermediate attributes Z). Further, human thinking is often framed as causal. Causal DAGs have allowed to formalize human reasoning in a ML-readable manner (Schölkopf et al., 2021).

Causal DAGs are able to graphically represent a worldview on a given fairness context, to highlight the (structural) assumptions, and to formalize the potential bias in a dataset (Pearl & Mackenzie, 2018). Causal DAGs have motivated the rise of causal fairness metrics (Makhlouf et al., 2020; Carey & Wu, 2022), including *total fairness* (Zhang & Bareinboim, 2018), *path-specific fairness* (Zhang et al., 2017), and *counterfactual fairness* (Kusner et al., 2017). Compared with the fairness notions based on correlation, causality-based fairness notions and methods include additional knowledge of the causal structure of the problem. This knowledge often reveals the mechanism of data generation, which helps comprehend and interpret the influence of sensitive attributes on the output of a decision process. This additional auxiliary causal knowledge, e.g., is often the basis for moving from testing unfairness to testing discrimination (Álvarez & Ruggieri, 2023). A common limitation is defining a causal DAG, which requires an agreement on its existence and, in turn, structure. It is not a straightforward task, but it also forces practitioners to state otherwise implicit assumptions about the data and encourages discussions among stakeholders (Kusner et al., 2017; Álvarez & Ruggieri, 2023).

Overall, while approaches for causal discovery from data can be adopted, specifically in the context of fairness (Binkyte-Sadauskiene et al., 2022), they definitively need to be complemented with domain expert knowledge—but, with no guarantee of an unanimous agreement among experts (Rahmattalabi & Xiang, 2022). Moreover, a number of assumptions are typically made which might not be met in practice, such as sufficiency (all causes are known), and faithfulness (the graph completely characterizes the conditional independences among features) (Spirtes et al., 2000). Further, causal fairness metrics may suffer from the identifiability problem (Makhlouf et al., 2022), namely the impossibility to compute them from observational data only.

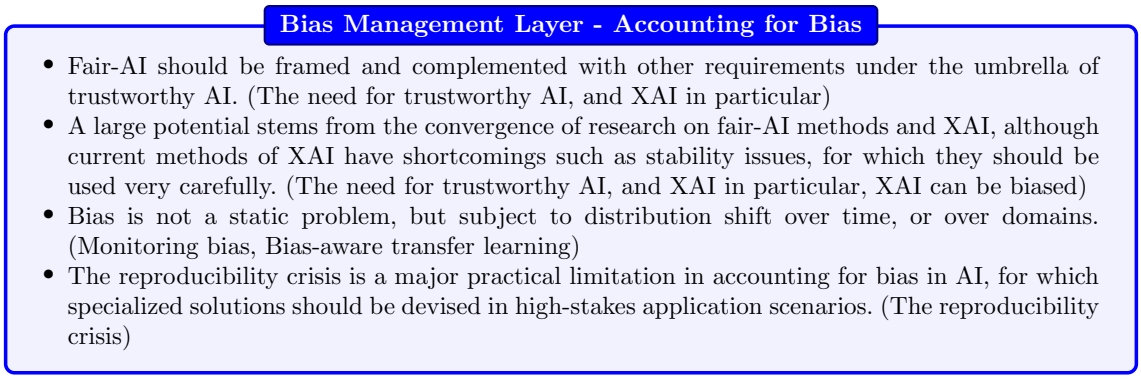

Fig. 5 Bias management layer—accounting for bias: challenges, policy advices, and best practices. In parenthesis, references to the subsections where they are discussed

Finally, the use of causal DAGs in fairness has not been free of criticism (e.g., Kasirzadeh and Smart (2021)). Arguments against the manipulability of the sensitive features, e.g., race, in counterfactual reasoning have been raised (Kohler-Hausmann, 2019; Hu & Kohler-Hausmann, 2020). These works argue that it is difficult, if not impossible, to disentangle the causal effects of the sensitive attributes on and from the other attributes in a meaningful way.

Knowledge-informed AI models

Relying exclusively on raw data for a given task is often not sufficient. Primarily, models trained on raw data fail to capture the nuances found in the less-represented segments of the data distribution (Mallen et al., 2023), which often correspond to underprivileged communities. While using external knowledge sources to compensate these inequalities holds promise (Lobo et al., 2023), this objective is not central to current knowledge-informed approaches. Typically, external sources support the so-called *knowledge-intensive tasks*, which are those tasks requiring a significant amount of real-world knowledge (e.g., fact verification) (Petroni et al., 2021). External knowledge sources are then used to update the model, provide higher interpretability, and enhance the reliability of its predictions (Asai et al., 2023). Other possible applications where informing predictions can be useful are based on using a combination of sources to enhance the generalizability of a model (Chiril et al., 2022). Particularly, leveraging data to improve performance outside the training distribution for a specific AI task. On issues closely related to discrimination, the integration of additional data and knowledge sources is gaining presence in the development of social-aware ML models (Wiegand et al., 2022). Such models are tailored to fill the gaps of individuals or groups with limited access to technology or who experience discriminatory representation, to frame AI systems within the specific social contexts in which they are applied.

The non-i.i.d. case: bias in unsupervised learning and graph-mining

The majority of traditional fair-AI metrics and methods are developed based on the independent and identically distributed (i.i.d.) data assumption: every instance in a dataset is drawn independently from a same statistical distribution. However, many real-world problems include graph-structured (network) data reflecting the connection between subjects, and such connections are not independent nor random—for instance, people connect due to similarity, local proximity, or common interests (Aiello et al., 2012). The studies centered on i.i.d. data are unable to reflect the bias exhibited by the relational information (i.e., the topology) in graph data. Fairness in graph mining can be non-trivial and it has exclusive backgrounds, taxonomies, and fulfilling techniques. Overviews papers by Dong et al. (2023); Chhabra et al. (2021); Choudhary et al. (2022), categorize a few of the current challenges and urgent needs in the field that we agree with. They include: (1) formulating (individual and group) fairness notions according to different types of biases and corresponding harms; (2) balancing model utility and algorithmic fairness; (3) explanation of bias in graph-based methods; and (4) enhancing the robustness of algorithms especially in cases of biased human annotations or malicious attacks. Harms of bias in the context of graphs, and in particular social networks, may go beyond discrimination, and include segregation (Baroni & Ruggieri, 2018; Ferrara et al., 2022), polarization (Tölle & Trier, 2023), filter bubbles (Pariser, 2011), and censorship (Aceto and Pescapè 2015). We see an urgent need for expanding the fair-AI research on the non-i.i.d. cases in the future.

Accounting for bias

In this section, we consider two technical aspects of accounting for bias, which complement the legal discussion of Sect. "AI fairness beyond metrics:transparency and accountabilityof AI systems": monitoring and explaining

bias. We claim the need for trustworthy AI as an holistic approach beyond fairness and bias issues. We warn, however, about the limitations of the young research field of XAI. Also, we discuss bias issues in tasks related to monitoring, including transferring AI models from a domain to another, and in reproducing evaluation scenarios. A summary of the challenges, policy advices, and best practices in accounting for bias is reported in Fig. 5.

The need for trustworthy AI, and XAI in particular

We think that the use of fair-AI methods should be complemented with design, development, and verification practices that are commonly summarized under the umbrella of *trustworthy AI* (Kaur et al., 2023). Such practices include: human agency and oversight, accountability, explainability, robustness and safety, privacy, diversity, reproducibility, and societal and environmental well-being. The research on the interplay between bias and those other non-functional requirements has been developing at different speeds. We refer to surveys on human-centered algorithmic fairness (Wu & Liu, 2022) (see also Sections 3.3.2), differential privacy and fairness (Fioretto et al., 2022), fairness and diversity constraints in ranking (Zehlike et al., 2023), trust and fairness (Knowles et al., 2022), and fairness and robustness (Lee et al., 2021a). A large potential stems from the convergence of fairness and XAI (Balkir et al., 2022; Rawal et al., 2022). XAI methods for model inspection, such as variable importance, can be used to test the influence/independence of protected attributes on a model's output (Grabowicz et al., 2022). Adding explanations to an AI system's output can increase users' trust and fairness perception (Tal et al., 2022) and ultimately control for the exercise of power (Lazar, 2022). In particular, local explanation methods that describe why a specific output was produced (*factual explanation*) and what could have changed the output (*counterfactual explanation*) can help to identify reasons of discriminatory decisions (Manerba & Guidotti, 2021) and to support actionable recourse (Karimi et al., 2023). XAI methods that aim to answer causal questions are referred to as causal interpretable models (Moraffah et al., 2020; Ganguly et al., 2023). Results of the NoBIAS project have considered desiderata for XAI in general, based on symbolic logic reasoning (State, 2022), and for the specific domain of central banking (Mougan et al., 2021). Different user profiles require a different level of explanations as well as different ways of integration to create a human-aligned conversational explanation system (Dazeley et al., 2021). Alarmingly, human evaluation is not the norm in the XAI field: considering the case of counterfactual explanations, Keane et al. (2021) found that only 21% of the approaches are validated with human subject experiments. For a summary of recent empirical findings and user studies in XAI research, see Vainio-Pekka et al. (2023); Rong et al. (2024). Moreover, the critical survey of Deck et al. (2023) points out a misalignment between fairness desiderata and the actual capabilities of the state-of-the-art in XAI.

XAI can be biased

Decision-making processes that affect individuals' rights and freedoms often require explanation (Kroll et al., 2017) (recall Sect. "AI fairness beyond metrics: transparency and accountability of AI systems"). While XAI methods offer a (non-exhaustive) way to hold AI systems accountable (Doshi-Velez et al., 2017), there are a number of limitations of current state-of-the-art that need to be acknowledged, and that should caution us from using these methods blindly. These limitations partly stem from the fact that research in XAI is relatively young (Confalonieri et al., 2021). A major problem is that when explaining black box models, multiple explanations are possible, possibly leading to disagreement about the reasons for the model's output (Krishna et al., 2022). Most prominently, post-hoc explainability methods, which typically rely on a surrogate interpretable model of a black box, are not guaranteed to be stable nor faithful to the underlying black box (Ghassemi et al., 2021). Possible gaps in faithfulness w.r.t. different sub-populations results then in potential biases also in the explanations, as shown for LIME and SHAP in Balagopalan et al. (2022). In an adversarial setting, i.e. a setting with different interests of the party explaining the ML model and the party receiving the explanation, this might allow for (intentionally) misleading explanations (Bordt et al., 2022). In line with this, other scholars argue that highly faithful explanations might not be desirable from a business perspective, and thus only carefully adopted, specifically regarding possible conflicts with Intellectual Property Rights (IPRs) and the potential to "game" the system (Barocas et al., 2020).

We highlight a few further issues of XAI. While there is a pool of explanation methods to pick from, most of them focus on classification tasks (and not, e.g., on unsupervised problems), and on tabular, image and text data (and not, e.g., on time series data). Being able to use an explainable AI method then implies that the problem might need to be adapted to the methods currently available, leading to possible losses of information and lower prediction accuracy (State et al., 2022). Also, interpreting explainability methods requires significant amounts of domain knowledge regarding the application context; a lack of such knowledge might render the explanations meaningless to (lay) end-users. Integration can be either achieved by involving the respective experts into the evaluation (see Sect. "Multi-stakeholder participatory design"), or by directly integrating it via symbolic approaches (Calegari et al., 2020), or knowledge-informed AI methods (see Sect. "Knowledge-informed AI models"). Beyond solving the technical issues of explainability methods as outlined above, there is also the need to adopt a holistic

perspective towards XAI, such as making sure that development teams are diverse, integrating all involved stakeholders into the design process (see Sect. "Multi-stakeholder participatory design"), evaluating XAI methods in context (see Sect. "The reproducibility crisis"), etc. Further, it might be worth investigating XAI methods and values embedded into these systems from other perspectives, such as that of historically marginalized groups (see Sect. "Intersectionality"). More research towards this is needed, and we point out emerging work such as State and Fahimi (2023), investigating explanations from a feminist perspective.

Monitoring bias

Model monitoring aims to evaluate model performance metrics, also w.r.t. bias and fairness, once the model has been deployed (Kenthapadi et al., 2022; Barrainkua et al., 2022). A common assumption in traditional batch ML is that bias is a static problem. This assumption is unrealistic for the many domains that have underlying *distribution shift* over time (Quiñonero-Candela et al., 2009). Subsequently, the problem of bias needs to be studied in continual (a.k.a. lifelong) learning scenarios (Lange et al., 2022), where AI models are continuously adapted to changing data. Another problem is the occurrence of *feedback loops* (Pagan et al., 2023; EU Agency for Fundamental Rights, 2022) (see also the notion of *performative predictions* (Perdomo et al., 2020)) which occur when the outputs of AI models subsequently affect the inputs to downstream systems. These vicious cycles can perpetuate unfairness even if static fair-AI models are used (Liu et al., 2018).

We distinguish two main categories of model monitoring. Supervised monitoring approaches rely on the availability of labelled data to compare the model's predictions against a set of ground truth labels. By evaluating the model's performance on this labelled data, we can identify performance deviations or biases that may have emerged during deployment. The NoBIAS project has contributed to this research by proposing approaches that use XAI methods for model monitoring and fairness auditing (Mougan and Nielsen, 2023; Mougan et al., 2022). However, labelled data may be available with an excessive delay (Lange et al., 2022), e.g., the data about defaults in loan repayment used to evaluate model's predictive accuracy. In some cases, labelled data may not be available at all, e.g., sensitive personal attributes to compare model's fairness across social groups may not be collectable due to data protection law (see Sect. "EU data protection law and non-discrimination law"). This is the case of the second category of model monitoring, namely unsupervised monitoring. Estimating the performance and fairness of AI models in the absence of labelled data is a very challenging task with impossibility theorems delimiting the work (Garg et al., 2022; Zhang et al., 2021; Fang et al., 2022).

Bias-aware transfer learning

It is common practice to adapt an upstream "pre-trained" model to a downstream task creating a downstream "target" model. Biases tend to be propagated when fine-tuning the source models to the downstream tasks (Salman et al., 2022). This propagation of biases is known as "bias transfer" (Steed et al., 2022). While bias transfer is a well-defined concept, it has mostly been explored within the context of NLP (Ladhak et al., 2023; Feng et al., 2023; Jin et al., 2021) except for (Salman et al., 2022) in the area of computer vision. Furthermore, since upstream bias mitigation is known to reduce bias transfer to the target models (Jin et al., 2021), we raise awareness about it as an effective bias mitigation step and encourage more research on its potential. Recent work by Álvarez et al. (2023) on decision tree classifiers under transfer learning, for instance, shows that incorporating partial knowledge from the target population (i.e., that on which the pre-trained model is to be deployed upon) when training the model can increase model performance and reduce the risk of unfair classifications. To some extent, this is again an instance of the knowledge-informed AI approach of Sect. "Knowledge-informed AI models".

The reproducibility crisis

The evaluation of AI models should replicate the operational scenario where the model will be deployed as closely as possible. Similarly, the auditing of AI models should replicate the operational scenario where the system has been deployed. Sometimes, "stress test" scenarios are also considered to assess the impact of improper usage of the AI models—this is the case of high-risk applications in the AI Act (see Sect. "The option not to use AI"). However, the lack of good documentation on AI development and bias management processes, including definitions, software, and datasets (see Sect. "Living with bias by documenting it"), are key factors affecting evaluation and reproducibility, giving raise to the *reproducibility crisis* (Gundersen, 2020). For instance, an issue has been raised about the arbitrariness of predictions of ML models trained across different samples (Cooper et al., 2023), showing that most fairness classification benchmarks are close-to-fair when taking into account such an arbitrariness. We see reproducibility as a major practical limitation in accounting for bias in AI, for which specialized solutions should be devised based on specific application scenarios. As an example, in the high-risk domain of credit scoring, the European Banking Authority[24] provides detailed guidelines and discussion papers including the monitoring of bias in ML models.

[24] https://www.eba.europa.eu/regulation-and-policy/credit-risk.

Conclusions

Many concerns about the risks and harms of bias in AI have been motivating the fast growing multidisciplinary research on fair-AI.

First, we have concisely summarized topics in policies and best practices, thus providing to researchers and practitioners pointers to inventories, guidelines and survey papers. On the one side of the spectrum of possible actions to prevent those risks and harms, there is the option not to use AI. On the other side of the spectrum, there is the option to address the origins of AI harms at societal level. In between the two options, there are methods for documenting bias, techniques for mitigating bias, and approaches for auditing AI systems.

Second, we have contributed to the ongoing fair-AI discussion with additional challenges, policy advice, and best practices that resulted from the execution of the NoBIAS project. We argue that these are, although deemed relevant, not sufficiently developed nor acknowledged in the literature. These are summarized in Figs. 2, 3, 4, 5, with in parenthesis the reference to the section of the paper where they are discussed in. While we do not claim for their completeness, we hope that those advices and best practices will contribute to the conventional wisdom in research and practice of managing bias and fairness in AI.

Acknowledgements This work has received funding from the European Union's Horizon 2020 research and innovation program under Marie Sklodowska-Curie Actions (Grant Agreement Number 860630) for the project "NoBIAS—Artificial Intelligence without Bias". 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.

Funding Open access funding provided by Università di Pisa within the CRUI-CARE Agreement.

Declarations

Conflict of interest We have no conflict of interest to disclose.

References

Aceto, G., & Pescapè, A. (2015). Internet censorship detection: A survey. *Computer Networks, 83*, 381–421.

Afzal, S., C., R., Kesarwani, M., et al. (2021). Data readiness report. In *SMDS. IEEE*, pp. 42–51

Aiello, L. M., Barrat, A., & Schifanella, R., et al. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web, 6*(2), 1–33.

Alkhatib, A. (2021). To live in their utopia: Why algorithmic systems create absurd outcomes. In: *CHI. ACM*, pp. 95:1–9

Almada, M. (2021). Automated decision-making as a data protection issue. Available at SSRN 3817472

Altman, A. (2020). Discrimination. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* Stanford University.

Álvarez, J.M., & Ruggieri, S. (2023). Counterfactual situation testing: Uncovering discrimination under fairness given the difference. In: *EAAMO. ACM*, pp. 2:1–11

Álvarez, J.M., Scott, K.M., & Berendt, B., et al. (2023). Domain adaptive decision trees: Implications for accuracy and fairness. In: *FAccT. ACM*, pp. 423–433

Alves, G., Bernier, F., Couceiro, M., et al. (2023). Survey on fairness notions and related tensions. *EURO Journal on Decision Processes, 11*, 100033.

Anisetti, M., Ardagna, C. A., Bena, N., et al. (2023). Rethinking certification for trustworthy machine-learning-based applications. *IEEE Internet Computing, 27*(6), 22–28.

Article 29 Data Protection Working Party. (2018). Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679 (wp251rev.01). https://ec.europa.eu/newsroom/article29/items/612053

Asai, A., Min, S., & Zhong, Z., et al. (2023). Retrieval-based language models and applications. In: ACL (tutorial). Association for Computational Linguistics, pp. 41–46

Ashurst, C., & Weller, A. (2023). Fairness without demographic data: A survey of approaches. In: EAAMO. ACM, pp 14:1–14

Asplund, J., Eslami, M., & Sundaram, H., et al. (2020). Auditing race and gender discrimination in online housing markets. In: ICWSM. AAAI Press, pp. 24–35

Balagopalan, A., Zhang, H., & Hamidieh, K., et al. (2022). The road to explainability is paved with bias: Measuring the fairness of explanations. In: FAccT. ACM, pp. 1194–1206

Balayn, A., & Gürses, S. (2021). *Beyond debiasing: Regulating AI and its inequalities*. European Digital Rights (EDRi): Tech. rep.

Balayn, A., Lofi, C., & Houben, G. (2021). Managing bias and unfairness in data for decision support: A survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal, 30*(5), 739–768.

Balayn, A., Yurrita, M., & Yang, J., et al. (2023). Fairness toolkits, a checkbox culture?" On the factors that fragment developer practices in handling algorithmic harms. In: AIES. ACM, pp. 482–495

Balkir, E., Kiritchenko, S., Nejadgholi, I., et al. (2022). Challenges in applying explainability methods to improve the fairness of NLP models. CoRR abs/2206.03945

Bao, M., Zhou, A., Zottola, S., et al. (2021). It's compaslicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. In: NeurIPS Datasets and Benchmarks

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*, 671–732.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org, http://www.fairmlbook.org

Barocas, S., Selbst, A.D., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In: FAT*. ACM, pp. 80–89

Baroni, A., & Ruggieri, S. (2018). Segregation discovery in a social network of companies. *Journal of Intelligent Information Systems, 51*(1), 71–96.

Barrainkua, A., Gordaliza, P., & Lozano, J.A., et al. (2022). A survey on preserving fairness guarantees in changing environments. CoRR abs/2211.07530

Bathaee, Y. (2018). The Artificial Intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology, 31*(2), 889–938.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604.

Berk, R., Heidari, H., Jabbari, S., et al. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50*(1), 3–44.

Beutel, A., Chen, J., & Doshi, T., et al. (2019). Putting fairness principles into practice: Challenges, metrics, and improvements. In: AIES. ACM, pp. 453–459

Bias (2023) Merriam-Webster.com Dictionary. Merriam-Webster, Inc.

Binkyte-Sadauskiene, R., Makhlouf, K., & Pinzón, C., et al. (2022). Causal discovery for fairness. CoRR abs/2206.06685

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research, 81*, 149–159.

Binns, R. (2018). What can political philosophy teach us about algorithmic fairness? *IEEE Security & Privacy, 16*(3), 73–80.

Binns, R. (2020). On the apparent conflict between individual and group fairness. In: FAT*. ACM, pp. 514–524

Binns, R., Adams-Prassl, J., & Kelly-Lyth, A. (2023). Legal taxonomies of machine bias: Revisiting direct discrimination. In: FAccT. ACM, pp. 1850–1858

Birhane, A., & Prabhu, V.U. (2021). Large image datasets: A pyrrhic win for computer vision? In: WACV. IEEE, pp. 1536–1546

Birhane, A., Prabhu, V.U., Kahembwe, E. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. CoRR abs/2110.01963

Blodgett, S.L., Barocas, S., & III, H.D., et al. (2020). Language (technology) is power: A critical survey of bias in NLP. In: ACL. Association for Computational Linguistics, pp. 5454–5476

Bordt, S., Finck, M., & Raidl, E., et al. (2022). Post-hoc explanations fail to achieve their purpose in adversarial contexts. In: FAccT. ACM, pp. 891–905

Budig, M. J., & England, P. (2001). The wage penalty for motherhood. *American Sociological Review, 66*(2), 204–225.

Buijsman, S .(2023). Navigating fairness measures and trade-offs. AI and Ethics

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 77–91.

Buyl, M., Bie, T.D. (2024). Inherent limitations of AI fairness. Commun ACM p to appear

Cabitza, F., Campagner, A., & Basile, V. (2023). Toward a perspectivist turn in ground truthing for predictive computing. In: AAAI. AAAI Press, pp. 6860–6868

Calegari, R., Ciatto, G., & Omicini, A. (2020). On the integration of symbolic and sub-symbolic techniques for XAI: A survey. *Intelligenza Artificiale, 14*(1), 7–32.

Camilleri, H., Ashurst, C., & Jaisankar, N., et al .(2023). Media coverage of predictive policing: Bias, police engagement, and the future of transparency. In: EAAMO. ACM, pp. 28:1–28:19

Carey, A. N., & Wu, X. (2022). The causal fairness field guide: Perspectives from social and formal sciences. *Frontiers Big Data, 5*, 892837.

Carey, A. N., & Wu, X. (2023). The statistical fairness field guide: Perspectives from social and formal sciences. *AI Ethics, 3*(1), 1–23.

Castelnovo, A., Crupi, R., Greco, G., et al. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports, 12*(1), 4209.

Castelnovo, A., Inverardi, N., & Nanino, G., et al. (2023). Fair enough? A map of the current limitations of the requirements to have fair algorithms. CoRR abs/2311.12435

Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. ACM Comput Surv p to appear

Caton, S., Malisetty, S., & Haas, C. (2022). Impact of imputation strategies on fairness in machine learning. *Journal of Artificial Intelligence Research, 74*, 1011–1035.

Chen, J., Dong, H., Wang, X., et al. (2023). Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems, 41*(3), 1–39.

Chen, R. J., Wang, J. J., Williamson, D. F. K., et al. (2023). Algorithmic fairness in Artificial Intelligence for medicine and healthcare. *Nature Biomedical Engineering, 7*(6), 719–742.

Chen, Z., Zhang, JM., & Hort, M., et al. (2022). Fairness testing: A comprehensive survey and analysis of trends. CoRR abs/2207.10223

Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially responsible AI algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research, 71*, 1137–1181.

Chhabra, A., Masalkovaite, K., & Mohapatra, P. (2021). An overview of fairness in clustering. *IEEE Access, 9*, 130698–130720.

Chiril, P., Pamungkas, E. W., Benamara, F., et al. (2022). Emotionally informed hate speech detection: A multi-target perspective. *Cognitive Computation, 14*(1), 322–352.

Chmielinski, KS., Newman, S., Taylor, M., et al. (2022). The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in Artificial Intelligence. CoRR abs/2201.03954

Choudhary, M., Laclau, C., Largeron, C. (2022). A survey on fairness for machine learning on graphs. CoRR abs/2205.05396

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.

Cobbe, J., Lee, MSA., & Singh, J. (2021). Reviewable automated decision-making: A framework for accountable algorithmic systems. In: FAccT. ACM, pp. 598–609

Confalonieri, R., Coba, L., Wagner, B., et al. (2021). A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11*(1), e1391.

Cooper, AF., Lee, K., Barocas, S., et al. (2023). Is my prediction arbitrary? Measuring self-consistency in fair classification. CoRR abs/2301.11562

Corbett-Davies, S., Pierson, E., Feller, A., et al. (2017). Algorithmic decision making and the cost of fairness. In: KDD. ACM, pp. 797–806

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Coston, A., Guha, N., & Ouyang, D., et al. (2021). Leveraging administrative data for bias audits: Assessing disparate coverage with mobility data for COVID-19 policy. In: FAccT. ACM, pp. 173–184

Council of the European Union (2000a) Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. Official Journal of the European Communities L 180. http://data.europa.eu/eli/dir/2000/43/oj

Council of the European Union (2000b) Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation. Official Journal of the European Communities L 303. http://data.europa.eu/eli/dir/2000/78/oj

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. Official Journal of the European Union L 373. http://data.europa.eu/eli/dir/2004/113/oj
Cummings, R., Gupta, V., & Kimpara, D., et al. (2019). On the compatibility of privacy and fairness. In: UMAP (Adjunct Publication). ACM, pp. 309–315
Czarnowska, P., Vyas, Y., & Shah, K. (2021). Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics, 9*, 1249–1267.
Danks, D., & London, A.J. (2017). Algorithmic bias in autonomous systems. In: IJCAI. ijcai.org, pp. 4691–4697
Dazeley, R., Vamplew, P., Foale, C., et al. (2021). Levels of explainable Artificial Intelligence for human-aligned conversational explanations. *Artificial Intelligence, 299*, 103525.
Deck, L., Schoeffer, J., & De-Arteaga, M., et al. (2023). A critical survey on fairness benefits of XAI. CoRR abs/2310.13007
Demartini, G., Roitero, K., & Mizzaro, S. (2023). Data bias management. *Communication ACM, 67*(1), 28–32.
Denton, E., Hanna, A., & Amironesei, R., et al. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data Society*. https://doi.org/10.1177/20539517211035955
D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT press.
Dong, Y., Ma, J., Chen, C., et al. (2023). Fairness in graph mining: A survey. IEEE Transactions on Knowledge and Data Engineering, pp. 1–22
Doshi-Velez, F., Kortz, M., & Budish, R., et al. (2017). Accountability of AI under the law: The role of explanation. CoRR abs/1711.01134
Draude, C., Klumbyte, G., Lücking, P., et al. (2019). Situated algorithms a sociotechnical systemic approach to bias. *Online Information Review, 44*(2), 325–342.
Dwork, C., Ilvento, C. (2019). Fairness under composition. In: ITCS, LIPIcs, vol 124. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 1–33
Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a right to an explanation is probably not the remedy you are looking for. *Tech Rev, 16*, 18.
Ehrhardt, A., Biernacki, C., Vandewalle, V., et al. (2021). Reject inference methods in credit scoring. *Journal of Applied Statistics, 48*, 2734–2754.
EU Agency for Fundamental Rights (2022) Bias in algorithms: Artificial intelligence and discrimination. Publications Office of the European Union, https://data.europa.eu/doi/10.2811/25847
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. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206
European Commission, Directorate-General for Justice and Consumers, & Gerards J, et al. (2021). Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law. Publications Office, https://data.europa.eu/doi/10.2838/544956
European Court of Justice. (2016). Parris v trinity college Dublin and others. (Case C-443/15)
European Parliament, 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 (recast). Official Journal of the European Union L 204. http://data.europa.eu/eli/dir/2006/54/oj
European Parliament, Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 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). Official Journal of the European Union L 119. http://data.europa.eu/eli/reg/2016/679/oj
European Parliament, Directorate-General for Parliamentary Research Services, & Beriain M., et al. (2022). Auditing the quality of datasets used in algorithmic decision-making systems. https://data.europa.eu/doi/10.2861/98930
European Union. (2000). Charter of Fundamental Rights of the European Union. Official Journal of the European Union C 364. http://data.europa.eu/eli/treaty/char_2012/oj
Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., et al. (2022). A survey on bias in visual datasets. *Computer Vision and Image Understanding, 223*, 103552.
Fabris, A., Messina, S., & Silvello, G., et al. (2022). Algorithmic fairness datasets: The story so far. *Data Mining and Knowledge Discovery*, *36*, 2074–2152
Fabris, A., Baranowska, N., & Dennis, M.J., et al. (2023). Fairness and bias in algorithmic hiring. CoRR abs/2309.13933
Fang, Z., Li, Y., & Lu, J., et al. (2022). Is out-of-distribution detection learnable? In: NeurIPS
Feffer, M., Skirpan, M., & Lipton, Z., et al. (2023). From preference elicitation to participatory ML: A critical survey & guidelines for future research. In: AIES. ACM, pp. 38–48
Feng, S., Park, CY., & Liu, Y., et al. (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In: ACL (1). Association for Computational Linguistics, pp. 11737–11762
Ferrara, A., Noboa, LE., & Karimi, F., et al. (2022). Link recommendations: Their impact on network structure and minorities. In: WebSci. ACM, pp. 228–238
Fiesler, C., Garrett, N., Beard, N. (2020). What do we teach when we teach tech ethics?: A syllabi analysis. In: SIGCSE. ACM, pp. 289–295
Fioretto, F., Tran, C., & Hentenryck, PV, et al. (2022). Differential privacy and fairness in decisions and learning tasks: A survey. In: IJCAI. ijcai.org, pp. 5470–5477
Foster, S. R. (2004). Causation in antidiscrimination law: Beyond intent versus impact. *Houston Law Review, 41*(5), 1469–1548
Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM, 64*(4), 136–143.
Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems, 14*(3), 330–347.
Future of Privacy Forum. (2017). Unfairness by algorithm: Distilling the harms of automated decision-making, https://fpf.org/blog/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/
Gajane, P., Saxena, A., Tavakol, M., et al. (2022). Survey on fair reinforcement learning: Theory and practice. CoRR abs/2205.10032
Galhotra, S., Shanmugam, K., & Sattigeri, P, et al. (2022). Causal feature selection for algorithmic fairness. In: SIGMOD Conference. ACM, pp. 276–285
Gallegos, IO., Rossi, RA., Barrow, J., et al. (2023). Bias and fairness in large language models: A survey. CoRR abs/2309.00770
Ganguly, N., Fazlija, D., & Badar, M., et al. (2023). A review of the role of causality in developing trustworthy AI systems. CoRR abs/2302.06975
Garg, S., Balakrishnan, S., Lipton, ZC., et al. (2022). Leveraging unlabeled data to predict out-of-distribution performance. In: ICLR. OpenReview.net
Garibay, Ö. Ö., et al. (2023). Six human-centered Artificial Intelligence grand challenges. *International Journal of Human-Computer Interaction, 39*(3), 391–437.
Gebru, T., Morgenstern, J., Vecchione, B., et al. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92.
Gellert, R., Vries, K.D., de Hert, P., et al. (2013). A comparative analysis of anti-discrimination and data protection legislations. In: Discrimination and Privacy in the Information Society, Studies in Applied Philosophy, Epistemology and Rational Ethics, vol 3. Springer, pp. 61–89

Gerards, J., & Zuiderveen Borgesius, F. J. (2022). Protected grounds and the system of non-discrimination law in the context of algorithmic decision-making and Artificial Intelligence. *Colorado Technology Law Journal, 20*, 1.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable Artificial Intelligence in health care. *Lancet Digit Health, 3*(11), e745–e750.

Gitelman, L. (2013). *Raw data is an oxymoron*. MIT Press.

Goel, N., Amayuelas, A., Deshpande, A., et al. (2021). The importance of modeling data missingness in algorithmic fairness: A causal perspective. In: AAAI. AAAI Press, pp. 7564–7573

Gohar, U., Cheng, L. (2023). A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. In: IJCAI. ijcai.org, pp. 6619–6627.

Grabowicz, P.A., Perello, N., & Mishra, A. (2022). Marrying fairness and explainability in supervised learning. In: FAccT. ACM, pp. 1905–1916.

Green, B., & Hu, L. (2018). The myth in the methodology: Towards a recontextualization of fairness in machine learning. In: Debates@ICML, https://econcs.seas.harvard.edu/files/econcs/files/green_icml18.pdf

Grimes, D. A., & Schulz, K. F. (2002). Bias and causal associations in observational research. *Lancet, 359*, 248–252.

Guidotti, R., Monreale, A., Ruggieri, S., et al. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42.

Gundersen, O. E. (2020). The reproducibility crisis is real. *AI Magazine, 41*(3), 103–106.

Gururangan, S., Card, D., Dreier, S.K., et al. (2022). Whose language counts as high quality? Measuring language ideologies in text data selection. In: EMNLP. Association for Computational Linguistics, pp. 2562–2580

Hacker, P. (2018). Teaching fairness to Artificial Intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review, 55*(4), 1.

Haeri, M.A., Zweig, K.A. (2020). The crucial role of sensitive attributes in fair classification. In: SSCI. IEEE, pp. 2993–3002

Harris, C., Halevy, M., Howard, A.M., et al. (2022). Exploring the role of grammar and word choice in bias toward African American English (AAE) in hate speech classification. In: FAccT. ACM, pp. 789–798

Haselton, M.G., Nettle, D., Andrews, P.W. (2005). The evolution of cognitive bias. In: Zalta EN (Eds.) Handbook of Evolutionary Psychology. John Wiley & Sons Inc., pp. 724–746

Hellström, T., Dignum, V., Bensch, S. (2020). Bias in machine learning - what is it good for? In: NeHuAI@ECAI, CEUR Workshop Proceedings, vol 2659. CEUR-WS.org, pp. 3–10

Hendrickx, K., Perini, L., der Plas, D.V., et al. (2021). Machine learning with a reject option: A survey. CoRR http://arxiv.org/abs/2107.11277

Henin, C., & Métayer, D. L. (2022). Beyond explainability: Justifiability and contestability of algorithmic decision systems. *AI Society, 37*(4), 1397–1410.

Hertweck, C., Heitz, C., & Loi, M. (2021). On the moral justification of statistical parity. In: FAccT. ACM, pp. 747–757

Hillman, T. (2011). The inscription, translation and re-inscription of technology for mathematical learning. *Technology, Knowledge and Learning, 16*(2), 103.

Hort, M., Chen, Z., Zhang, J.M., et al. (2022). Bias mitigation for machine learning classifiers: A comprehensive survey. CoRR http://arxiv.org/abs/2207.07068

Hsee, C. K., & Li, X. (2022). A framing effect in the judgment of discrimination. *Proceedings of the National Academy of Sciences, 119*(47), e2205988119.

Hu, L., & Kohler-Hausmann, I. (2020). What's sex got to do with machine learning? In: FAT*. ACM, p. 513

Hughes, M. M. (2011). Intersectionality, quotas, and minority women's political representation worldwide. *American Political Science Review, 105*(3), 604–620.

Hutchinson, B., Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. In: FAT. ACM, pp. 49–58

Hutchinson, B., Smart, A., Hanna, A., et al. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In: FAccT. ACM, pp. 560–575

ISO/IEC. (2021). ISO/IEC TR 24027:2021 - Information Technology - Artificial Intelligence (AI) - Bias in AI systems and AI-aided decision making. https://www.iso.org/standard/77607.html

Jacobs, A.Z. (2021). Measurement and fairness. In: FAccT. ACM, pp. 375–385

Jaton, F. (2020). *The Constitution of Algorithms*. Ground-Truthing, Programming, Formulating: Inside technology, The MIT Press

Ji, D., Smyth, P., Steyvers, M. (2020). Can I trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. In: NeurIPS

Ji, J., Qiu, T., Chen, B., et al. (2023). AI alignment: A comprehensive survey. CoRR http://arxiv.org/abs/2310.19852

Jin, X., Barbieri, F., Kennedy, B., et al. (2021). On transferability of bias mitigation effects in language model fine-tuning. In: NAACL-HLT. Association for Computational Linguistics, pp. 3770–3783

Jo, E.S., Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In: FAT*. ACM, pp. 306–316

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Jorgensen, M., Richert, H., Black, E., et al. (2023). Not so fair: The impact of presumably fair machine learning models. In: AIES. ACM, pp. 297–311

Kahneman, D. (2011). *Thinking*. Farrar, Straus and Giroux: Fast and Slow

Kahneman, D., Sibony, O., Sunstein, C. (2021). Noise: A Flaw in Human Judgment. William Collins

Kamiran, F., Calders, T. (2009). Classifying without discriminating. In: International conference on computer, control and communication. IEEE, pp. 1–6

Kamishima, T., Akaho. S., Asoh, H., et al. (2012). Fairness-aware classifier with prejudice remover regularizer. In: ECML/PKDD (2), LNCS, vol 7524. Springer, pp. 35–50

Karimi, A., Barthe, G., Schölkopf, B., et al. (2023). A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys, 55*(5), 1–29.

Kasirzadeh, A., & Smart, A. (2021). The use and misuse of counterfactuals in ethical machine learning. In: FAccT. ACM, pp. 228–236

Katell, M.A., Young, M., Herman, B., et al. (2019). An algorithmic equity toolkit for technology audits by community advocates and activists. CoRR http://arxiv.org/abs/1912.02943

Kaur, D., Uslu, S., Rittichier, K. J., et al. (2023). Trustworthy artificial intelligence: A review. *ACM Computing Surveys, 55*(2), 1–38.

Kazim, E., Koshiyama, A. S., Hilliard, A., et al. (2021). Systematizing audit in algorithmic recruitment. *Journal of Intelligence, 9*(3), 46.

Keane, M.T., Kenny, E.M., Delaney, E., et al. (2021). If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. In: IJCAI. ijcai.org, pp. 4466–4474

Kenthapadi, K., Lakkaraju, H., Natarajan, P., et al. (2022). Model monitoring in practice: Lessons learned and open challenges. In: KDD. ACM, pp. 4800–4801

Kiviat, B. (2019). The art of deciding with data: evidence from how employers translate credit reports into hiring decisions. *Socio-Economic Review, 17*(2), 283–309.

Kleinberg, J.M., Mullainathan, S., Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In: ITCS,

LIPIcs, vol 67. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 43:1–23
Knowles, B., Richards, J.T., Kroeger, F. (2022). The many facets of trust in AI: Formalizing the relation between trust and fairness, accountability, and transparency. CoRR http://arxiv.org/abs/2208.00681
Koch, G., & Kinder-Kurlanda, K. (2020). Source criticism of data platform logics on the internet. *Historical Social Research, 45*(3), 270–287.
Kohler-Hausmann, I. (2019). Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern University Law Review, 113*(5), 1163–1227.
Koshiyama, A., Kazim, E., Treleaven, P., et al. (2021). Towards algorithm auditing: A survey on managing legal, ethical and technological risks of AI, ML and associated algorithms. Available at SSRN: https://doi.org/10.2139/ssrn.3778998
Kraft, A., & Usbeck, R. (2022). The lifecycle of "facts": A survey of social bias in knowledge graphs. In: AACL/IJCNLP (1). Association for Computational Linguistics, pp. 639–652
Krishna, S., Han, T., Gu, A., et al. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. CoRR http://arxiv.org/abs/2202.01602
Kroll, J. A., Huey, J., Barocas, S., et al. (2017). Accountable algorithms. *U of Penn Law Review, 165*, 633–705.
Kulynych, B., Overdorf, R., Troncoso, C., et al. (2020). Pots: protective optimization technologies. In: FAT*. ACM, pp. 177–188
Kusner, M.J., Loftus, J.R., Russell, C., et al. (2017). Counterfactual fairness. In: NIPS, pp. 4066–4076
Ladhak, F., Durmus, E., Suzgun, M., et al. (2023). When do pre-training biases propagate to downstream tasks? A case study in text summarization. In: EACL. Association for Computational Linguistics, pp. 3198–3211
Lakkaraju, H., Kleinberg, J.M., Leskovec, J., et al. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In: KDD. ACM, pp. 275–284
Lange, M. D., Aljundi, R., Masana, M., et al. (2022). A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on Pattern Analysis and Machine Intelligence, 44*(7), 3366–3385.
Lazar, S. (2022). Legitimacy, authority, and the political value of explanations. CoRR abs/2208.08628
Lebovitz, S., Levina, N, & Lifshitz-Assaf, H. (2021). Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. MIS Q 45(3)
Lee, J., Roh, Y., Song, H., et al. (2021a). Machine learning robustness, fairness, and their convergence. In: KDD. ACM, pp. 4046–4047
Lee, M. S. A., & Floridi, L. (2021). Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds and Machines, 31*(1), 165–191.
Lee, M.S.A., Singh, J. (2021). The landscape and gaps in open source fairness toolkits. In: CHI. ACM, pp. 1–13
Lee, M. S. A., Floridi, L., & Singh, J. (2021). Formalising trade-offs beyond algorithmic fairness: Lessons from ethical philosophy and welfare economics. *AI Ethics, 1*(4), 529–544.
Li, H., Vincent, N., Chancellor, S., et al. (2023). The dimensions of data labor: A road map for researchers, activists, and policymakers to empower data producers. In: FAccT. ACM, pp. 1151–1161
Lin, C. K., & Jackson, S. J. (2023). From bias to repair: Error as a site of collaboration and negotiation in applied data science work. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), 1–32.
Lin, Z. J., Jung, J., Goel, S., et al. (2020). The limits of human predictions of recidivism. *Science Advances, 6*(7), 0652.
Liu, L. T., Dean, S., Rolf, E., et al. (2018). Delayed impact of fair machine learning. *International Conference on Machine Learning, 80*, 3156–3164.
Lobo, P. R., Daga, E., Alani, H., et al. (2023). Semantic web technologies and bias in Artificial Intelligence: A systematic literature review. *Semantic Web, 14*(4), 745–770.
Lopez, P. (2019). Reinforcing intersectional inequality via the AMS algorithm in Austria. In: Proc. of the STS Conference. Verlag der Technischen Universität Graz, pp. 289–309
Lowry, S., & Macpherson, G. (1986). A blot on the profession. *British Medical Journal, 296*(6623), 657–658.
Madaio, M., Egede, L., Subramonyam, H., et al. (2022). Assessing the fairness of AI systems: AI practitioners processes, challenges, and needs for support. *Proceedings of the ACM on Human-Computer Interaction, 6*, 1–26.
Majumder, S., Chakraborty, J., Bai, G. R., et al. (2023). Fair enough: Searching for sufficient measures of fairness. *ACM Transactions on Software Engineering and Methodology, 32*(6), 1–22.
Makhlouf, K., Zhioua, S., & Palamidessi, C. (2020). Survey on causal-based machine learning fairness notions. CoRR abs/2010.09553
Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021). Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management, 58*(5), 102642.
Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021). On the applicability of machine learning fairness notions. *SIGKDD Explorations Newsletter, 23*(1), 14–23.
Makhlouf, K., Zhioua, S., & Palamidessi, C. (2022). Identifiability of causal-based fairness notions: A state of the art. CoRR abs/2203.05900
Mallen, A., Asai, A., Zhong, V., et al. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In: ACL (1). Association for Computational Linguistics, pp. 9802–9822
Manerba, M.M., & Guidotti, R. (2021). Fairshades: Fairness auditing via explainability in abusive language detection systems. In: CogMI. IEEE, pp. 34–43
Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics, 160*(4), 835–850.
Mehrabi, N., Morstatter, F., Saxena, N., et al. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35.
Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence, 5*, 100152.
Mendoza, I., & Bygrave, L. A. (2017). *The right not to be subject to automated decisions based on profiling* (pp. 77–98). EU Internet Law: Regulation and Enforcement
Menon, A. K., & Williamson, R. C. (2018). The cost of fairness in binary classification. *Proceedings of Machine Learning Research, 81*, 107–118.
Metcalf, J., Moss, E., Watkins, E.A., et al. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. In: FAccT. ACM, pp. 735–746
Miceli, M., Posada, J., & Yang, T. (2022). Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction, 6*, 1–14.
Miceli, M., Yang, T., Garcia, A. A., et al. (2022). Documenting data production processes: A participatory approach for data work. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1–34.
Miller, A.P. (2018). Want less-biased decisions? Use algorithms. Harvard Business Review
Minh, D., Wang, H. X., Li, Y. F., et al. (2022). Explainable Artificial Intelligence: A comprehensive review. *Artificial Intelligence Review, 55*(5), 3503–3568.
Minow, M. (2021). Equality vs. Equity. *American Journal of Law and Equality, 1*, 167–193.
Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model cards for model reporting. In: FAT. ACM, pp. 220–229

Mitchell, S., Potash, E., Barocas, S., et al. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*, 141–163.

Mittelstadt, B.D., Wachter, S., & Russell, C. (2023). The unfairness of fair machine learning: Levelling down and strict egalitarianism by default. CoRR abs/2302.02404

Mökander, J. (2023). Auditing of AI: legal, ethical and technical approaches. *Digital Society, 2*(3), 49.

Moraffah, R., Karami, M., Guo, R., et al. (2020). Causal interpretability for machine learning - problems, methods and evaluation. *SIGKDD Explorations Newsletter, 22*(1), 18–33.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., et al. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review, 56*(4), 3005–3054.

Mougan, C., & Nielsen, D.S. (2023). Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap. In: AAAI. AAAI Press, pp. 15037–15045

Mougan, C., Kanellos, G., & Gottron, T. (2021). Desiderata for explainable AI in statistical production systems of the european central bank. In: PKDD/ECML Workshops (1), Communications in Computer and Information Science, vol 1524. Springer, pp. 575–590

Mougan, C., Broelemann, K., Kasneci, G., et al. (2022). Explanation shift: Detecting distribution shifts on tabular data via the explanation space. In: NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications

Mougan, C., Álvarez, J.M., Ruggieri, S., et al. (2023). Fairness implications of encoding protected categorical attributes. In: AIES. ACM, pp. 454–465

Mulligan, D. K., Kroll, J. A., Kohli, N., et al. (2019). This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction, 3*, 1–36.

Nader, L. (1972). Up the anthropologist: Perspectives gained from studying up. Tech. Rep. ED065375, ERIC, https://eric.ed.gov/?id=ED065375

Nogueira, A. R., Pugnana, A., Ruggieri, S., et al. (2022). Methods and tools for causal discovery and causal inference. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12*(2), e1449.

Ntoutsi, E., Fafalios, P., Gadiraju, U., et al. (2020). Bias in data-driven Artificial Intelligence systems - An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1356.

van Nuenen, T., Such, J. M., & Coté, M. (2022). Intersectional experiences of unfair treatment caused by automated computational systems. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1–30.

Olteanu, A., Castillo, C., Diaz, F., et al. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers Big Data, 2*, 13.

Organizers Of QueerinAI, et al. (2023). Queer in AI: A case study in community-led participatory AI. In: FAccT. ACM, pp. 1882–1895

Ovalle, A., Subramonian, A., Gautam, V., et al. (2023). Factoring the matrix of domination: A critical review and reimagination of intersectionality in AI fairness. In: AIES. ACM, pp. 496–511

Pagan, N., Baumann, J., Elokda, E., et al. (2023). A classification of feedback loops and their relation to biases in automated decision-making systems. CoRR abs/2305.06055

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press.

Parmar, M., Mishra, S., Geva, M., et al. (2023). Don't blame the annotator: Bias already starts in the annotation instructions. In: EACL. Association for Computational Linguistics, pp. 1771–1781

Passi, S., Barocas, S. (2019). Problem formulation and fairness. In: FAT. ACM, pp. 39–48

Pearl, J. (2009). Causality: models, reasoning and inference, Second Edition. Cambridge University Press

Pearl, J., Mackenzie, D. (2018). The book of why: The new science of cause and effect. Basic Books

Pedreschi, D., Ruggieri, S., Turini, F. (2008). Discrimination-aware data mining. In: KDD. ACM, pp. 560–568

Pedreschi, D., Ruggieri, S., Turini, F. (2012). A study of top-k measures for discrimination discovery. In: SAC. ACM, pp. 126–131

Peng, K., Mathur, A., Narayanan, A. (2021). Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In: NeurIPS Datasets and Benchmarks

Percy, C., Dragicevic, S., Sarkar, S., et al. (2021). Accountability in AI: from principles to industry-specific accreditation. *AI Communications, 34*(3), 181–196.

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., et al. (2020). Performative prediction. *Proceedings of Machine Learning Research, 119*, 7599–7609.

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys, 55*(3), 1–44.

Petroni, F., Piktus, A., Fan. A., et al. (2021). KILT: a benchmark for knowledge intensive language tasks. In: NAACL-HLT. Association for Computational Linguistics, pp. 2523–2544

Pleiss, G., Raghavan, M., Wu, F., et al. (2017). On fairness and calibration. In: NIPS, pp. 5680–5689

Pruss, D. (2023). Ghosting the machine: Judicial resistance to a recidivism risk assessment instrument. In: FAccT. ACM, pp. 312–323

Quiñonero-Candela, J., Sugiyama, M., Lawrence, N. D., et al. (2009). *Dataset shift in machine learning*. MIT Press.

Quy, T. L., Roy, A., Iosifidis, V., et al. (2022). A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12*(3), e1452.

Rahmattalabi, A., Xiang, A. (2022). Promises and challenges of causality for ethical machine learning. CoRR abs/2201.10683

Raji, I. D., Yang, J. (2019). ABOUT ML: annotation and benchmarking on understanding and transparency of machine learning lifecycles. CoRR abs/1912.06166

Raji, I. D., Smart, A., White, R. N., et al. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: FAT*. ACM, pp. 33–44

Raji, I. D., Bender, E. M. et al. (2021a). AI and the everything in the whole wide world benchmark. In: NeurIPS Datasets and Benchmarks

Raji, I. D., Scheuerman, M. K., Amironesei, R. (2021b). You can't sit with us: Exclusionary pedagogy in AI ethics education. In: FAccT. ACM, pp. 515–525

Rawal, A., McCoy, J., Rawat, D. B., et al. (2022). Recent advances in trustworthy explainable Artificial Intelligence: Status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence, 3*(6), 852–866.

Räz T (2021) Group fairness: Independence revisited. In: FAccT. ACM, pp. 129–137

Richardson, B., Gilbert, J. E. (2021). A framework for fairness: A systematic review of existing fair AI solutions. CoRR abs/2112.05700

Rismani, S., Moon, A. (2023). What does it mean to be a responsible AI practitioner: An ontology of roles and skills. In: AIES. ACM, pp. 584–595

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review, 29*(5), 582–638.

Rong, Y., Leemann, T., Nguyen, T., et al. (2024). Towards human-centered explainable AI: A survey of user studies for model explanations. IEEE Trans Pattern Anal Mach Intell p to appear

Rovatsos, M., Mittelstadt, B., Koene, A. (2019). Landscape Summary: Bias In Algorithmic Decision-Making: What is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it? UK Government

Roy, A., Horstmann, J., Ntoutsi, E. (2023). Multi-dimensional discrimination in law and machine learning - A comparative overview. In: FAccT. ACM, pp. 89–100

Ruggieri, S., Álvarez, J. M., Pugnana, A., et al. (2023). Can we trust fair-AI? In: AAAI. AAAI Press, pp. 15421–15430

Sadiq, S. W., Aryani, A., Demartini, G., et al. (2022). Information resilience: the nexus of responsible and agile approaches to information use. *The VLDB Journal, 31*(5), 1059–1084.

Saha, D., Schumann, C., McElfresh, D. C., et al. (2020). Measuring non-expert comprehension of machine learning fairness metrics. *Proceedings of Machine Learning Research, 119*, 8377–8387.

Salman, H., Jain, S., Ilyas, A., et al. (2022). When does bias transfer in transfer learning? CoRR abs/2207.02842

Saltz, J. S., Skirpan, M., Fiesler, C., et al. (2019). Integrating ethics within machine learning courses. *ACM Transactions on Computing Education, 19*(4), 1–26.

Scantamburlo, T. (2021). Non-empirical problems in fair machine learning. *Ethics and Information Technology, 23*(4), 703–712.

Schölkopf, B., Locatello, F., Bauer, S., et al. (2021). Toward causal representation learning. Proc IEEE *109*(5), 612–634.

Schwartz, R., Vassilev, A., Greene, K., et al. (2022). Towards a standard for identifying and managing bias in Artificial Intelligence. Tech. Rep. 1270, NIST Special Publication

Scott, K. M., Wang, S. M., Miceli, M., et al. (2022). Algorithmic tools in public employment services: Towards a jobseeker-centric perspective. In: FAccT. ACM, pp. 2138–2148

Seaver, N. (2017). Algorithms as culture. Some tactics for the ethnography of algorithmic systems. *Big Data & Society, 4*(2), 2053951717738104.

Shahbazi, N., Lin, Y., Asudeh, A., et al. (2023). Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys*. https://doi.org/10.1145/3588433

Shahriar, S., Allana, S., Hazratifard, S. M., et al. (2023). A survey of privacy risks and mitigation strategies in the Artificial Intelligence life cycle. *IEEE Access, 11*, 61829–61854.

Shelby, R., Rismani, S., Henne, K., et al. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In: AIES. ACM, pp. 723–741

Silberzahn, R., & Uhlmann, E. L. (2015). Crowdsourced research: Many hands make tight work. *Nature, 526*, 189–191.

Smirnov, I., Lemmerich, F., & Strohmaier, M. (2021). Quota-based debiasing can decrease representation of the most under-represented groups. *Royal Society Open Science, 8*(9), 210821.

Spirtes, P., & Zhang, K. (2016). Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics, 3*(3), 1–38.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction, and Search* (2nd ed.). Adaptive computation and machine learning: MIT Press.

Srivastava, M., Heidari, H., Krause, A. (2019). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In: KDD. ACM, pp. 2459–2468

State, L. (2022). Constructing meaningful explanations: Logic-based approaches. In: AIES. ACM, p. 916

State, L., Fahimi, M. (2023). Careful explanations: A feminist perspective on XAI. In: EWAF, CEUR Workshop Proceedings, vol 3442. CEUR-WS.org

State, L., Salat, H., Rubrichi, S., et al. (2022). Explainability in practice: Estimating electrification rates from mobile phone data in senegal. CoRR abs/2211.06277

Steed, R., Panda, S., Kobren, A., et al. (2022). Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In: ACL (1). Association for Computational Linguistics, pp. 3524–3542

Stoyanovich, J., Abiteboul, S., Howe, B., et al. (2022). Responsible data management. *Communications of the ACM, 65*(6), 64–74.

Suresh, H., Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In: EAAMO. ACM, pp. 17:1–17:9

Sylolypavan, A., Sleeman, D. H., Wu, H., et al. (2023). The impact of inconsistent human annotations on AI driven clinical decision making. *NPJ Digit Medicine, 6*, 26.

Szczekocka, E., Tarnec, C., Pieczerak, J. (2022). Standardization on bias in Artificial Intelligence as industry support. In: Big Data. IEEE, pp. 5090–5099

Tal, A. S., Kuflik, T., & Kliger, D. (2022). Fairness, explainability and in-between: Understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology, 24*(1), 2.

Tal, E. (2023). Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare. In: AIES. ACM, pp. 312–321

Tang, Z., Zhang, J., & Zhang, K. (2023). What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Computing Surveys, 55*, 1–37.

Tizpaz-Niari, S., Kumar, A., Tan, G., et al. (2022). Fairness-aware configuration of machine learning libraries. In: ICSE. ACM, pp. 909–920

Tolan, S. (2019). Fair and unbiased algorithmic decision making: Current state and future challenges. arXiv preprint arXiv:1901.04730

Tölle, L., Trier, M. (2023). Polarization in online social networks: A review of mechanisms and dimensions. In: ECIS

Turri, V., Dzombak, R. (2023). Why we need to know more: Exploring the state of AI incident documentation practices. In: AIES. ACM, pp. 576–583

Vainio-Pekka, H., otse Agbese MO, Jantunen M, et al. (2023). The role of explainable AI in the research field of AI ethics. *ACM Transactions on Interactive Intelligent Systems, 13*(4), 1.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society, 4*(2), 2053951717743530.

Vedder, A., & Naudts, L. (2017). Accountability for the use of algorithms in a big data environment. *International Review of Law, Computers & Technology, 31*(2), 206–224.

Verma, S., Rubin, J. (2018). Fairness definitions explained. In: FairWare@ICSE. ACM, pp. 1–7

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law. *W Va L Rev, 123*(3), 735–790.

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review, 41*, 105567.

Wagstaff, K. (2012). Machine learning that matters. In: ICML. icml.cc/Omnipress

Wan, M., Zha, D., Liu, N., et al. (2023). In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data, 17*(3), 1–27.

Wang, A., Kapoor, S., Barocas, S., et al. (2023). Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. In: FAccT. ACM, p. 626

Wei, S., & Niethammer, M. (2022). The fairness-accuracy Pareto front. *Statistical Analysis and Data Mining, 15*(3), 287–302.

Weinberg, L. (2022). Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ML fairness approaches. *Journal of Artificial Intelligence Research, 74*, 75–109.

Wick, M. L., Panda, S., Tristan, J. (2019). Unlocking fairness: A trade-off revisited. In: NeurIPS, pp. 8780–8789

Wiegand, M., Eder, E., Ruppenhofer, J. (2022). Identifying implicitly abusive remarks about identity groups using a linguistically informed approach. In: NAACL-HLT. ACL, pp. 5600–5612

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In: FAT*. ACM, pp. 1–18

Wu, D., Liu, J. (2022). Involve humans in algorithmic fairness issue: A systematic review. In: iConference (1), LNCS, vol 13192. Springer, pp. 161–176

Xenidis, R. (2018). Multiple discrimination in EU anti-discrimination law: Towards redressing complex inequality? In: Belavusau, U., Henrard, K. (Eds.) EU anti-discrimination law beyond gender. Hart Publishing, pp. 41–74

Xenidis, R. (2020). Tuning EU equality law to algorithmic discrimination: Three pathways to resilience. *Maastricht Journal of European and Comparative Law, 27*(6), 736–758.

Xenidis, R., & Senden, L., et al. (2020). EU non-discrimination law in the era of Artificial Intelligence: Mapping the challenges of algorithmic discrimination. In U. Bernitz (Ed.), *General principles of EU law and the EU digital order* (pp. 151–182). Kluwer Law International.

Xu, W. (2019). Toward human-centered AI: A perspective from human-computer interaction. *Interactions, 26*(4), 42–46.

Zajko, M. (2022). Artificial Intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass, 16*(3), e12962.

Zehlike, M., Yang, K., & Stoyanovich, J. (2023). Fairness in ranking, part I: Score-based ranking. *ACM Computing Surveys, 55*, 1–36.

Zhang, J., Bareinboim, E. (2018). Fairness in decision-making - the causal explanation formula. In: AAAI. AAAI Press, pp. 2037–2045

Zhang, L., Wu, Y., Wu, X. (2017). A causal framework for discovering and removing direct and indirect discrimination. In: IJCAI. ijcai.org, pp. 3929–3935

Zhang, L. H., Goldstein, M., & Ranganath, R. (2021). Understanding failures in out-of-distribution detection with deep generative models. *Proceedings of Machine Learning Research, 139*, 12427–12436.

Zhang, Z., Wang, S., & Meng, G. (2023). A review on pre-processing methods for fairness in machine learning. *Advances in natural computation, Fuzzy Systems and Knowledge Discovery* (pp. 1185–1191). Springer.

Ziems, C., Chen, J., Harris, C., et al. (2022). VALUE: understanding dialect disparity in NLU. In: ACL (1). Association for Computational Linguistics, pp. 3701–3720

Ziewitz, M. (2016). Governing algorithms. Myth, mess, and methods. *Science Technology Human Values, 41*(1), 3–16.

Zliobaite, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery, 31*(4), 1060–1089.

Zliobaite, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law, 24*(2), 183–201.

Zuiderveen Borgesius, F. J. (2020). Strengthening legal protection against discrimination by algorithms and Artificial Intelligence. *The International Journal of Human Rights, 24*(10), 1572–1593.

Authors and Affiliations

Jose M. Alvarez[1,2] · Alejandra Bringas Colmenarejo[3] · Alaa Elobaid[4,5] · Simone Fabbrizzi[4,6,7] · Miriam Fahimi[8] · Antonio Ferrara[9,10] · Siamak Ghodsi[5,6] · Carlos Mougan[3] · Ioanna Papageorgiou[6] · Paula Reyero[11] · Mayra Russo[6] · Kristen M. Scott[12] · Laura State[1,2] · Xuan Zhao[13] · Salvatore Ruggieri[2]

✉ Jose M. Alvarez
jose.alvarez@sns.it

✉ Salvatore Ruggieri
salvatore.ruggieri@unipi.it

Alejandra Bringas Colmenarejo
Alejandra.Bringas-Colmenarejo@soton.ac.uk

Alaa Elobaid
elobaida@iti.gr

Simone Fabbrizzi
simone.fabbrizzi@unibz.it

Miriam Fahimi
miriam.fahimi@aau.at

Antonio Ferrara
antonio.ferrara@centai.eu

Siamak Ghodsi
ghodsi@l3s.de

Carlos Mougan
c.mougan@soton.ac.uk

Ioanna Papageorgiou
ioanna.papageorgiou@iri.uni-hannover.de

Paula Reyero
paula.reyero-lobo@open.ac.uk

Mayra Russo
mrusso@l3s.de

Kristen M. Scott
kristen.scott@kuleuven.be

Laura State
laura.state@di.unipi.it

Xuan Zhao
xuan.zhao@schufa.de

[1] Scuola Normale Superiore, Pisa, Italy

[2] University of Pisa, Pisa, Italy

[3] University of Southampton, Southampton, UK

[4] CERTH, Thessaloniki, Greece

[5] Free University of Berlin, Berlin, Germany

[6] Leibniz University Hannover, Hannover, Germany

[7] Free University of Bozen-Bolzano, Bolzano, Italy

[8] University of Klagenfurt, Klagenfurt, Austria

[9] GESIS - Leibniz Institute, Mannheim, Germany

[10] RWTH Aachen University, Aachen, Germany

[11] The Open University, Milton Keynes, UK

[12] KU Leuven, Leuven, Belgium

[13] SCHUFA Holding AG, Wiesbaden, Germany