Pfeiffer, Jella et al.

Algorithmic Fairness in AI

Business & Information Systems Engineering

Provided in Cooperation with:
Springer Nature

Suggested Citation: Pfeiffer, Jella et al. (2023) : Algorithmic Fairness in AI, Business & Information Systems Engineering, ISSN 1867-0202, Springer Fachmedien Wiesbaden GmbH, Wiesbaden, Vol. 65, Iss. 2, pp. 209-222, https://doi.org/10.1007/s12599-023-00787-x

This Version is available at:
https://hdl.handle.net/10419/311040

Bus Inf Syst Eng
https://doi.org/10.1007/s12599-023-00787-x

DISCUSSION

Algorithmic Fairness in AI

An Interdisciplinary View

Jella Pfeiffer · Julia Gutschow · Christian Haas · Florian Möslein · Oliver Maspfuhl · Frederik Borgers · Suzana Alpsancar

1 Motivation

Jella Pfeiffer, Julia Gutschow

In 2016, an investigative journalism group called ProPublica analyzed COMPAS, a recidivism prediction algorithm based on machine learning used in the U.S. criminal justice sector. This instrument assigns risk scores to defendants that are supposed to reflect how likely that person is to commit another crime upon release. The group found that the instrument was much more likely to falsely flag black defendants as high risk and less likely to falsely assess them to be low risk than it was the case for white defendants. ProPublica assessed this to be highly problematic as false decisions in this area of application can have a major impact on the defendants' lives, possibly affecting their prospects of early release, probationary conditions or the amount of bail posted (Angwin et al. 2016). This example from the criminal justice sector shows that discrimination is not only a problem of human but also of algorithmic decision-making. Algorithmic fairness is particularly interesting when considering machine learning algorithms because they typically learn from past data, which might already be biased. Furthermore, a machine learning algorithm that tends to make unfair decisions might lead to systematic discrimination because, once trained, the algorithm might decide for a large amount of future cases. As such AI algorithms are used in many contexts such as personalized advertising, recruiting, credit business, or pricing (Dastile et al. 2020; Lambrecht and Tucker 2019; Raghavan et al. 2020; Sweeney 2013), they can gravely impact the further development of peoples' lives both on the individual and on the societal level, e.g., by increasing the wealth gap, but also impact organizations, e.g., by violating equal opportunity policies (Kordzadeh and Ghasemaghaei 2022). It is, therefore, of utmost importance to not only ensure that AI systems do not discriminate systematically but, going one step further, to also understand them as a chance to mitigate potential unfairness stemming from human-based decision-making.

This discussion paper mainly draws from a symposium on algorithmic fairness that was held in March 2022 in line with the 100th annual conference of the German Academic Association of Business Research (VHB). The symposium was interdisciplinary with speakers from the fields of philosophy and ethics, business and information systems engineering, law, as well as practice representatives from

J. Pfeiffer (✉) · J. Gutschow
Justus Liebig University Gießen, Giessen, Germany
e-mail: jella.pfeiffer@wirtschaft.uni-giessen.de

J. Gutschow
e-mail: julia.j.gutschow@wirtschaft.uni-giessen.de

C. Haas
Vienna University of Economics and Business (WU), Vienna, Austria
e-mail: christian.haas@wu.ac.at

F. Möslein
Philipps-University Marburg, Marburg, Germany
e-mail: florian.moeslein@jura.uni-marburg.de

O. Maspfuhl
Deutsche Bank AG, Frankfurt, Germany
e-mail: oliver.maspfuhl@db.de

F. Borgers
UNIQA Insurance Group AG, Vienna, Austria
e-mail: frederik.borgers@uniqa.at

S. Alpsancar
Paderborn University, Paderborn, Germany
e-mail: suzana.alpsancar@uni-paderborn.de

Published online: 06 February 2023

the banking and the insurance sector. The discussion that ensued due to this plethora of perspectives consolidated the decision to retain the most interesting insights in writing.

The symposium yielded five core themes which are discussed in this paper from several perspectives. We think that an interdisciplinary approach like this is exceptionally important when addressing a topic that is of such high relevance for society, economy and governments. This paper therefore includes viewpoints from the research on business and information systems (Prof. Dr. Christian Haas), from law (Prof. Dr. Florian Möslein), from the banking industry (Dr. Oliver Maspfuhl) as well as the insurance industry (Dr. Frederik Borgers), and from philosophy and ethics (Jun.-Prof. Suzana Alpsancar).

In a first step, we tackle the persisting problem of defining fairness. Throughout the years, the research community has constructed many criteria of fairness (Mehrabi et al. 2021; Verma and Rubin 2018; Yona 2017). However, many of the criteria are mutually exclusive, making it necessary to evaluate on a case-by-case basis which ones should be used when developing AI systems. In some cases, a decision made by an AI system may be fair with respect to objective fairness criteria, but the affected person may still subjectively feel discriminated against. How do we deal with these situations? Can we simply object to this feeling?

Next, we explore differences between human and algorithmic decision-making. Often, decisions made by AI systems are assumed to be inherently more objective and unbiased than those formed by human decision-making as the first are based on data and at least not directly influenced by subliminal human prejudices. AI systems are equipped to make decisions more efficiently and consistently than human decision-makers can. But despite their illusion of neutrality, algorithmic decision-making systems, and particularly those using machine learning, often contain the same biases as human decision-making because they heavily rely on past data as input. When the input data is biased, future decisions of the algorithmic decision-making system may be as well. We, therefore, ask ourselves whether the implementation of AI leads to a reproduction of discrimination or whether it can also help to reduce discrimination. To what extent and in which application areas are AI systems a better fit than human decision-makers when it comes to making fair decisions?

As a third core theme, we investigate approaches to mitigate discrimination in AI systems. Pre-, in- and post-processing techniques intervene at different stages of the algorithmic decision-making process, with pre-processing techniques focusing on the training data, while in-processing techniques tackle the algorithm itself, and post-processing techniques consider the decision outcomes (Mehrabi et al. 2021). We discuss the benefits and drawbacks of the different approaches and explore whether the applicability of these techniques differs between contexts. Are there trade-offs between fairness improvements and the accuracy of the decisions made by the AI system?

The fourth core theme reflects upon the EU AI Act, an intended European law proposed by the European Commission aiming to regulate the AI market. It proposes a risk-based differentiation of AI systems that prohibits particularly harmful AI practices while setting legal requirements for AI systems that are assessed to be high-risk. In line with the intended regulation, AI systems classified as low-risk would only have to follow minor transparency obligations while those classified as minimal risk are permitted with no restriction (European Commission 2021). Here, we aim to examine how the draft regulation will shape the framework conditions for using AI in the long term and to what extent companies are already preparing for this now.

Finally, the fifth core theme is concerned with the long-term impact that AI will have in the future. Other consequences, such as data protection or cybercrime, may have to be more intensively evaluated when implementing AI solutions. Aiming to bridge the gap between theory and practice and across disciplines, this discussion paper aims to provide an outlook on further research and the next steps for the practice.

2 Insights from Information Systems Research

Christian Haas

Over the last 10 years, research into Algorithmic Fairness, or Fairness in (data-driven) decision making, has seen considerable attention in the information systems (IS) and computer science (CS) communities (among others, of course), largely due to the pervasive collection and use of data in everyday decision making (Corbett-Davies et al. 2017; Feuerriegel et al. 2020). Yet, the question of what discrimination and fairness is, and how it can be defined, has a long history starting with the U.S. Civil Rights acts in the 1960s. Specifically, the years after the introduction of Title VI and Title VII laws (prohibiting discrimination in employment) saw the emergence of fundamental concepts and definitions of fairness, many of which are still used today (Hutchinson and Mitchell 2019). A core focus of this early research was on fundamental questions: (i) What is fairness, and how can it be defined? (ii) Can we quantify, and thus measure, fairness in a decision process? (iii) How are different fairness definitions related to each other and can several definitions be achieved simultaneously?

This focus on a quantitative definition of fairness has led to over two dozen fairness definitions, yet we still see no

convergence towards a universal definition (even though some definitions are more frequently used than others). One particular challenge of this plethora of definitions is that many of which are effectively incompatible with each other (Mitchell et al. 2021). In other words, achieving a fair outcome according to one definition can mean that a fair outcome based on another definition is not possible. For instance, many fairness definitions compare the prediction of a decision process (using a score S) for different groups (A) to the actual outcome (Y). These group fairness measures can be simplified according to three main concepts of fair outcomes: independence, separation, and sufficiency (Barocas et al. 2018). Independence considers an outcome as fair if the acceptance rates are equal across groups (the score S needs to be independent from the group membership A). Separation, instead, compares error rates across groups (the prediction score S needs to be independent from the group membership A, conditional on the actual outcome Y). Finally, sufficiency considers the distribution of the actual outcome given a scoring rule of the decision process (the outcome distribution Y needs to be independent of the group membership, conditional on the score S). If, for example, the actual outcome (Y) and the group membership (A) are not independent, independence and sufficiency cannot hold simultaneously. In addition, in non-trivial settings, the independence of the outcome (Y) and the group membership (A) can also lead to the incompatibility of separation and sufficiency (Castelnovo et al. 2022).

An example of these incompatibilities, and the challenging conversations that arise when a specific fairness definition needs to be selected, is the previously mentioned criminal recidivism case and the COMPAS dataset. The decision algorithm predicts whether or not a person is likely to recommit another crime, given risk profile scores. One group, ProPublica, concluded that the algorithm is unfair because of large differences in the false positive and false negative rates between white and black defendants, i.e., the percentage of defendants incorrectly flagged as likely or unlikely to recommit a crime (Angwin et al. 2016). Specifically, the corresponding separation-related fairness metric, equality of odds, was not satisfied. In contrast, a second group highlighted the similar predictive parity of the predictions, a metric related to the sufficiency principle, and argued that this is a more useful definition of fairness in this case (Flores et al. 2016). Connecting this to the previous concepts of fairness, the incompatibility of the considered fairness definitions resulted from a different true recidivism rate for the different groups (the outcome Y was not independent from the group membership A), in which case the two fairness definitions, one related to the separation principle, the other to the sufficiency principle, could not be achieved together (Chouldechova 2017).

Over the years, especially with the uptake of fairness research in the IS and CS communities, further questions were considered in addition to the definition of fairness itself (Mehrabi et al. 2021): (i) What is the impact of (specific) fairness definitions on other aspects of the decision process, such as decision quality/performance? (ii) Which strategies and adjustments to the decision process can be used to reduce unfairness and mitigate biases?

Algorithmic Fairness is often seen from the lens of a fairness versus performance trade-off. Specifically, adjusting the algorithm or decision process such that specific definitions of fairness can be achieved or improved can lead to a decrease of the accuracy of predictions (Chen et al. 2018; Menon and Williamson 2018). Yet, the impact on other aspects of the decision process, even alternative performance metrics, is less clear. For instance, while the general incompatibility of specific fairness definitions mentioned earlier is well established, these incompatibility results do not quantify the exact impact of enforcing one fairness measure over another, i.e., how one fairness measure changes at the cost of another. Here, IS research tries to provide more general frameworks to quantify the impact of achieving specific definitions of fairness on other performance metrics (and also other fairness definitions) of the decision process (Haas 2019). In addition, fairness considerations are increasingly examined in a wider decision context to measure the potential impact of enforcing fairness as compared to other aspects of the decision-making process. For example, implementing specific definitions of fairness can have an impact on the strategic behavior of companies. Fu et al. (2022) show switching from an independence-based fairness definition to a separation-based definition can lead to an underinvestment in the learning process for the underlying decision algorithm. This can then translate into outcomes that make both majority (advantaged) and minority (disadvantaged) groups (customers) worse off compared to the initial scenario.

Another stream of research in Algorithmic Fairness focuses on novel mitigation strategies to improve fairness (and avoid biases in the decisions). As mentioned before, the strategies tackle different stages of the decision-making process, i.e., either the data themselves (pre-processing), the algorithm or decision procedure (in-processing), or the predictions/decisions (post-processing). Especially the last 10 years have seen a substantial increase in the number of these bias mitigation approaches (Caton and Haas 2020). The majority of this work on bias mitigation strategies analyzes novel mitigation strategies against an unmitigated baseline, yet does not consider the effects of a potential combination of mitigation strategies across the decision process. For instance, instead of only transforming the data through a pre-processing approach, using the transformed data in a subsequent in- or post-processing strategy could

further improve the resulting fairness of the process outcomes. While such an ensemble of mitigation approaches could potentially yield additional benefits, comparing the dozens of potential mitigation strategies at any given stage of the decision process is practically impossible and current research lacks guidance into which mitigation strategies to use in which context.

Besides discussing the core challenges of incompatible fairness definitions and the lack of clear guidance for bias mitigation strategies, over the past years, IS research into Algorithmic Fairness has branched out to consider additional aspects. On the one hand, fairness considerations have been applied to specific scenarios such as hiring processes (Raghavan et al. 2020). On the other hand, researchers have begun to shift the focus from achieving fairness in a (conceptually) self-contained decision process to further aspects such as the consideration of the socio-technical environment in which the decision process is situated (Dolata et al. 2021). Data-driven decisions are not self-contained processes. Instead, they are parts of a larger environment and context that includes different actors and goals. While algorithmic decision-making has once been perceived as being more objective than human decisions due to its sole reliance on data, it is now well known that data frequently includes biases stemming from various sources (Mehrabi et al. 2021). For example, data used in a decision process can have a representation bias where certain minorities are not adequately represented, or it can have a selective labels bias where the observations stem from a human decision process and certain outcomes and variables were not observed (Kleinberg et al. 2017). Hence, finding a "fair" comparison of how data-driven decisions compare against human decisions is a separate research direction by itself. Finally, recent research increasingly considers fairness along with aspects of explainability and transparency in the more general context of human-AI decision-making (Alufaisan et al. 2021; Dodge et al. 2019; Shulner-Tal et al. 2022).

3 Legal and Normative Aspects

Florian Möslein

From a legal perspective, fairness plays a crucial role in different areas of law, and notions of fairness have been given remarkable academic attention. Due to its vagueness, however, the meanings and implications of the term vary considerably depending on the specific legal context. In contract law, for example, the fairness standard differs in the pre-contractual phase and within contractual relationships (Willett 2007). Generally, fair equality of opportunities counts among the core legal principles ever since John Rawls' groundbreaking article on "justice as fairness" (Rawls 1958). In the law and economics literature, the legal notion of fairness is often contrasted with the core economic concept of efficiency, thereby highlighting its defining role for the legal sphere (Kaplow and Shavell 2002). In legal discourse, the distinction between procedural fairness and substantive fairness is fundamental: Whereas the former concerns the process that leads to a decision or an agreement, the latter looks at its substance (similar to the outcome), e.g., at how rights and obligations are distributed (Allan 1998). Another important distinction draws the line between commutative and distributive fairness: "we identify 'commutative' as related to justice in exchange […] which is governed by the principle of equality, and which occurs between persons taken as individuals, while 'distributive' applies to the allocation of goods within a structure (a society, a firm etc.), which operates on the basis of proportionality" (Sadurski 2011, p. 94).

With respect to technology, the notion of fairness is often used in a rather unspecified sense but in fact relates to a very substantive, distributive idea of fairness: "A technological intervention to which the Fairness Principle applies is morally right only if it does not lead to unfair inequalities in society" (Peterson 2017, p. 168). From that viewpoint, the concept of fairness is closely linked to the principle of non-discrimination while procedural aspects lose all of their importance. Non-discrimination, in turn, is frequently used in legal provisions, not least because it provides a more specific yardstick than the concept of fairness. At the European level, for instance, various directives on equal treatment have been adopted in order to protect people from discrimination based on race, religion or belief, disability, age, gender or sexual orientation (Ellis and Watson 2012). Since non-discrimination rules are linked to such specific criteria, they only prevent unfairness if it results in a corresponding discrimination. On a more general level, the law does not prohibit any kind of behavior that may subjectively feel unfair: Whereas subjective fairness perceptions differ widely, legal provisions require objective standards that are as specific as possible in order to provide effective yardsticks. In a legal sense, fairness is therefore not "in the eye of the beholder" (Konow 2009).

Against the background of these deep and diverse conceptual foundations of fairness, it is difficult to specify what the term precisely means in relation to AI. Some indication is to be found in the so-called Ethics Guidelines for Trustworthy Artificial Intelligence that have been published by the High-Level Expert Group on Artificial Intelligence (Hleg AI 2018), an independent expert group that was set up by the European Commission. The Guidelines count fairness among the "four ethical

principles, rooted in fundamental rights, which must be respected in order to ensure that AI systems are developed, deployed and used in a trustworthy manner" (Hleg AI 2018, p. 12 ff.; see also Möslein and Horn 2021, p. 80 ff.). Moreover, they emphasize the many different interpretations of fairness and differentiate in particular between a substantive and a procedural dimension. As to the substantive dimension, the importance of ensuring equal and just distribution of both benefits and costs is stressed. By accentuating that AI should also ensure individuals and groups to be free from unfair bias, discrimination and stigmatisation, the Ethics Guidelines also illustrate that non-discrimination forms part of the more general concept of fairness (Hleg AI 2018, p. 12). The procedural dimension of fairness, on the other hand, is described so as to entail the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them (Hleg AI 2018, p. 13). More particularly, the Guidelines specify that the entity accountable for the decision must be identifiable and that the decision-making processes should be explicable. While the Ethical Guidelines thus elaborate in quite some detail what fairness implies, they are of an entirely voluntary nature: Stakeholders committed towards achieving trustworthy AI can opt to use these Guidelines as a method to operationalise their commitment (Hleg AI 2018, p. 5). Nonetheless, their fairness principles may well develop into a yardstick for AI systems because the Guidelines create a normative standard that enjoys the support of the European Commission as well as practical recognition. Non-compliance can therefore have substantial negative reputational effects (Möslein and Horn 2021, pp. 87–89). However, it does not result in any legal sanctions and the principles cannot be enforced before the courts or by public authorities. From a normative perspective, their nature is therefore fundamentally different from legal rules (Möslein 2022, p. 82 ff.).

At a more formal, legal level, rules on AI are emerging as well. In April 2021, the European Commission published a proposal for a respective regulation, the so-called AI Act (European Commission 2021). This proposal aims to establish harmonized rules for the placement on the market, the putting into service, and the use of artificial intelligence systems (Bomhard and Merkle 2021; Ebers et al. 2021; Veale and Borgesius 2021). In substance, it pursues a risk-based approach by establishing four different risk classes (Mahler 2022). Depending on this risk classification, the regulatory intensity increases, ranging from minimal, to medium, high, and unacceptable risk exposure (European Commission 2021, p. 3). In contrast with the Ethical Guidelines, it is striking that the AI Act does not even mention the fairness principle. Quite the contrary, the term "fair" is exclusively used with regard to the EU Charter of Fundamental Rights (CFR) which itself relies on fairness ideas when it establishes, for instance, the right to fair working conditions (Art. 31 CFR) or to a fair trial (Art. 47 CFR). As the use of AI with its specific characteristics like opacity, complexity, dependency on data, or autonomous behavior can adversely affect a number of fundamental rights enshrined in that Charter, the AI Act proposal seeks to ensure a high level of protection for these fundamental rights and aims to address various sources of risks through its risk-based approach (European Commission 2021, p. 11). In addition, the proposal also references the ideas of trustworthiness laid down in the Ethics Guidelines by aiming "to ensure the proper functioning of the single market by creating the conditions for the development and use of trustworthy artificial intelligence in the Union" (European Commission 2021, pp. 6 and 9). Whereas references to fairness are therefore of a relatively hidden nature, the AI Act proposal refers more explicitly to the more specific requirement of non-discrimination, at least in its explanatory memorandum and recitals. For example, Recital 15 sets out that AI technology can be "misused and provide novel and powerful tools for manipulative, exploitative and social control practices", and it stresses that such practices are particularly harmful and should be prohibited because they contradict, inter alia, the right to non-discrimination. In general, the prevention of discriminatory outcomes of AI systems is reflected in numerous parts of the AI Act (cf. also Recitals 17, 28, 35, 36, 37, 39, 44, 45, 47) and it significantly shaped the overall conceptual framework of the proposal (Ince 2021, p. 3). The proposal aims to supplement existing discrimination law (European Commission 2021, p. 4). The objective to prevent discriminatory outcomes had a decisive influence on the risk classification of the systems. For example, the enumerative list of systems that, according to Art. 5 of the proposal, should either be completely or partially prohibited contains the scoring of citizens for general purposes. These kinds of AI systems may lead to a detrimental treatment or even an exclusion of whole groups of people. They are therefore regarded as a violation of the right to non-discrimination, the right to equality, and even human dignity. Therefore, Art. 5 para.1 lit.c) prohibits the use of AI systems which are intended to establish a classification system for the trustworthiness of people, based on an evaluation of the social behavior by public authorities (so-called social scoring) (cf. Recital 17). With respect to the category of high-risk AI systems, the comprehensive list of obligations in Art. 9–15 AI-Act is also shaped by the idea to complement existing provisions on non-discrimination law by imposing various obligations to avert discrimination caused by algorithms, such as the requirement of a risk management system (Art.9) the obligations of transparency (Art.13), and human oversight (Art.14) (Recital 44; see also Veale and Borgesius 2021, p. 101 ff.; Townsend 2021,

p. 4). Moreover, Art. 10 requires the provider to ensure the quality of datasets by requiring the establishment of data governance and management procedures as well as introducing an obligation that the training, testing and validation datasets must be complete, error-free, and representative. Because the quality of the data is crucial to avoid biased outcomes of an AI system, this obligation highlights the intention of the Commission to prevent algorithm-based discrimination right from the origin of its emergence. Whereas the AI Act does not explicitly spell out any general fairness principles and, more generally, takes a rather instrumental and procedural approach to the regulation of artificial intelligence, this more specific aim to prevent discrimination is reflected in various parts of the proposal, in particular in the requirements for high-risk systems and in relation to the general risk classification.

4 Insights from the Banking Industry

Oliver Maspfuhl

4.1 AI in the Banking Industry – an Ethical Challenge

Quantitative methods of data analysis and modelling for risk assessment and forecasting – typically referred to as Machine Learning (ML) today – have been a part of the DNA of financial institutions for centuries. Their rebranding as Artificial Intelligence (AI) inspired by applications to computer vision or natural language understanding should not obscure this fact. In the banking industry, the introduction of Basel II was a booster towards data- and evidence-based decision-making, which made it inevitable for larger institutions to use statistical models for predicting and managing bank credit exposure and capital requirements. It is notable that, in contrast to many technical applications of AI for engineering purposes, financial applications were concerned with making decisions on human individuals since their inception, and, thus, were naturally confronted with ethical questions. The core challenge is dealing with individuals which – as opposed to machines or cars – can never be even approximately identical in a statistical sense.

4.2 What Characterises the Decision-Making of an AI System?

Although a fundamental distinction between "classical statistical models" and "modern AI systems" has been conjectured many times, there is no evidence of such a clear cut. Although very different in complexity, in practice, most AI applications are based on Machine Learning models [including classical ones like Generalized Linear Models (GLM)] and share the same basic characteristics (we restrict ourselves to supervised models for simplicity):

1. Their purpose is to provide the correct mapping between known inputs and unknown outputs when this relation is *not obvious*, *complex*, and *cannot be derived as a logical consequence of fundamental principles or assumptions*, but can, in principle, *be observed.*
2. The mapping is obtained by adapting *generic* mathematical structures with free parameters to known examples of input and output pairings so that the prediction is *most likely* to be correct for new inputs, but *not causally connected* to them.
3. The *result* of the application of the model is *deterministic* and, in principle, can be expressed as a (complex) mathematical formula. The empirical *correctness* of the predicted outcome, in contrast, is a *random (Bernoulli)* variable.

Particularly the last point often leads to confusion and needs more explanation:

3a. The result is obtained in two steps: At the core of the AI application, there is a Machine Learning model that maps inputs to a probability for each possible outcome. In a second step, the prediction is determined based on some defined decision rule, e.g., choosing the most likely outcome. The predicted probability and the decision are reached deterministically, but the decision is only correct with the predicted probability and thus, its *correctness* is stochastic.

These general settings have to be kept in mind when discussing whether the decision-making of AI systems is fair.

The classic use case discussed in the banking context – strongly boosted by Basel II – is the determination of creditworthiness and credit decisions by AI systems. Based on input values like income, length of client relationship, or other relevant criteria, an applicant for a loan will be assigned a so-called probability of default (PD) used for decisions concerning the granting and pricing of the loan – typically, this is no binary decision. Instead, the price of the loan will be adjusted according to the PD. The fairness of the price is obviously very relevant and can have a huge social impact. Notice, however, that the situation is somewhat more subtle as we have to assess the fairness of a *probability* here.

4.3 The Role of AI in the Fairness Debate

Ever since the discourse on fairness in AI has risen in relevancy, we observe an ongoing debate around the very

meaning of fairness. Trying to define fairness is a hopeless endeavour. In fact, it is *not* a problem that has emerged with the advent of AI systems. Making decisions under uncertainty and in the absence of clear evidence has always been the heavy burden of lawmakers and judges. The reason AI acts as a game changer is another: Being able to *automate* such decisions, based either on records, on known human decisions from the past, or known observed outcomes from the past, it becomes possible to considerably scale the amount of decisions without human intervention (in principle) and with a *deterministic* result (cf. point 3. above) that leaves no room for adaptation to individual circumstances once the set of relevant input data has been fixed in the model design phase. Obviously, that means that it is no longer sufficient to explain the reason for a decision in an individual case, as a jury would do at the announcement of their verdict in a trial, and which could be called an *individual* a posteriori explanation. Instead, as the model output is completely determined by the input values, the very logic of the decision needs to be explainable *universally and* a priori.

4.4 No Individual Fairness in AI

Unfortunately, this leads us into a vicious circle: Looking at our primordial principle (1.) above, we see that it is impossible to give such an explanation due to the very definition of a ML model: First, if the relationship of inputs (e.g., income, age, or region of residence) to the output (the (non-)default of the customer) were exactly known, that is, described by an exact structural formula representing a strict causality, there would be no need to use a data-driven model to come up with a prediction. Second, as stated in principle (3a), the model will just predict a probability. According to Popper's classical paradigm, a model can only be considered a valid explanation of reality if it clearly states how it can be falsified, e.g., which *individual* empirical observation will prove it wrong. However, a model predicting *probabilities* can only be proven wrong on an *ensemble of observations*. In summary, we see that fairness, in the framework of ML-based AI systems, is a concept for *groups of individuals,* not for individuals. As we established earlier that there are no identical human individuals, the key question starts to emerge: *Under which circumstances can human individuals be treated as a peer group*?

4.5 It all Boils Down to Transferring Group Properties to Individuals

Explanatory techniques that are useful to understanding Machine Learning models do not offer an explanation in the scientific sense but rather help accentuate the role individual inputs, also called features, play for the determination of the output in general and for individual predictions. This understanding is crucial. To improve the design of the model, the definition of peer groups (feature level sets) can be optimized to better represent individuals and lower the likelihood of unfair decisions, which may occur if an individual happens to be an outlier with respect to the average relationship represented by the model. It goes without saying that building ML-based AI models is therefore *not* a task for IT specialists and that AI, in this broader sense, is not to be considered a subset of computer science. It is a complex subject based on the mathematical modelling of data which is best accomplished by mixed teams of senior specialists with business, modelling, and IT backgrounds.

4.6 Discrimination Versus Non-discrimination

Thinking again of fairness in the sense of non-discrimination, we need to recall that the very purpose of Machine Learning models is to discriminate between input values that correspond to different output values, e.g., defaulting and non-defaulting loans. Non-discrimination may be easy to achieve by granting everyone the same conditions, but this would make the use of AI inefficient. Even worse, it could also be considered as violating the equally important principle of *equity*: A high-risk customer with no resources would get the same conditions as an individual with large savings, although both represent very different risks. Resulting losses might lead to the failure of the bank and inflict damages on its customers. Thus, *non-discrimination can be unfair,* too. In addition, granting loans to customers who default on their payback obligations will often result in a worsening of their situation. Note: There are cases where models are trained on human decisions from the past that may have been unfair, discriminating, or simply wrong. Here, we focus on cases where labels have been obtained by an objective process, e.g., real credit defaults.

4.7 Assessing Fairness a Posteriori

In this case, and in the light of the above arguments, it is useful to design any predictive model in such a way that it tries to make the best prediction given past evidence and to assess and ensure any fairness properties only a posteriori. The best practice for an a posteriori treatment is to define which features should be marked as sensitive in the sense that we do not want, by ethical principles, to get different model outcomes for input values differing *in those features only*. Practically, this means that, in the case that we do not want different loan prices for men and women, we would determine the price as always being the average of the model output with the sensitive features taking all possible

values. Notice that this is only possible if the sensitive attribute is known to the model. Otherwise, a potential discrimination is even impossible to detect.

4.8 Techniques from Classical Risk Management

In practice, one would typically use portfolio-weighted averages. This is an application of the well-known insurance principle of risk pooling, replacing highly variable individual risks with manageable portfolio averages. The primary aim of this risk management technique is not per se a fairer risk pricing, however, it does lead to a more targeted pricing of the risk and thus a more transparent and effective credit portfolio steering in line with regulators' aspirations. In response to the discussion around the fairness of ML models, individuals and their rights are now shifting into the focus of the design of credit risk models. Contemporary advances of these models are nonetheless well prepared to also ensure a maximum of individual fairness. However, there is a flip side: If more individual information is represented in the model, more personal data needs to be revealed, resulting in less solidarity among individuals.

4.9 Conclusion and Recommendation

1. Concluding from the above considerations, and in view of the experience gained over the last 15 years, it seems that the best strategy is to rely on the following principles for fair model design, irrespective of the type of model that is used:
2. Use real default data to avoid human decision bias. Make sure no population is underrepresented in the training data due to overly exclusive credit decisions.
3. Build the best model possible using all features that should be used, including sensitive ones, but excluding personal data (e.g., sexual orientation). The model should thoroughly include checks to see whether features can reasonably be generalized (e.g., a residence region might be an indication of current income, however, origin or sex should not be taken as a proxy for income since these attributes cannot be altered by the person).
4. Correct for any unwanted but evidence-supported discrimination via portfolio averages.
5. Make sure the features (or rather their common occurrence patterns) relevant for model decisions are made transparent to the individual and that they can be questioned and complemented by other evidence in the individual case.

Higher complexity in terms of structure or number of parameters will make those aims more ambitious. However, the interpretability of a model and its performance are not incompatible and, thus, do not have to be balanced in a "trade-off". They constitute two mutually supportive aspects to be improved simultaneously to reach a common optimum. There is no point in transparency for incorrect predictions and any model correctly reflecting reality must be plausible and understandable – however sophisticated it might need to be in order to represent a complex reality adequately.

5 Insights from the Insurance Industry

Frederik Borgers

Insurance, despite its private character, has a strong collective component. Think about mandatory insurance such as Motor Third Party Liability (MTPL), workmen's compensation, or the social role insurance plays in the case of natural catastrophes. This implies that every individual should be given fair access to protection by insurance and that not just regulators but also the industry should take any possible discrimination very seriously. In my contribution, I will focus on the risk of unfair pricing practices for motor insurance. MTPL is a homogeneous, mandatory product where market positions are mainly determined by pricing. This does not mean that possible discrimination is limited to pricing alone, but rather that access to a fair price and to the product itself is a first condition for the insurance market to function correctly.

Typically, the basis for a price calculation in motor insurance is a so-called risk model, which is a predictive model estimating the claim's cost per individual policy. For this purpose, a historical database of policies and claims, enriched with several external data sources, is used. For a long time, generalized linear models (GLM) were the industry standard, however, in recent years, AI techniques have gained popularity, often in combination with human influence or control. In a way, a price based purely on risk could be considered fair as each market segment would pay the premium they "deserve" based on their claim history as a group.

As insurers act in a competitive environment, their pricing, however, is not just based on risk alone. Typically, they will try to model demand using historical quotes and their conversion rate, which is the number of successful offers divided by the total number of offers. Using the combination of risk and conversion models then allows to create different scenarios where the central question turns around the preferred volume, i.e., the profit mix. The aim is to reach the "efficient border", meaning a status where, at a given volume, the profitability is maximized or vice versa. Specialized optimization algorithms are used to reach this efficient border. While all of this sounds like a very

sophisticated, data-driven approach, the use of sales discounts persists on the European insurance market until today, depending on the country and way of distribution. Based on this, three different types of price discrimination can be distinguished: (i) Risk-based discrimination, (ii) demand-based discrimination, and (iii) intermediary discrimination through sales discounts.

Risk-based Discrimination can occur due to the inclusion of discriminatory predictive variables in the risk model, assuming that these variables are used in the same way in the final tariff. There are various, relatively common examples with relevance for potential discrimination:

1. The EU has banned discrimination based on gender for the pricing of insurance products.
2. In Switzerland, it is common to use nationality or country of origin as a tariff factor, causing immigrants from non-EU countries, for instance, to pay significantly higher prices. In the EU itself, this practice is banned.
3. One of the most distinctive risk factors in motor insurance is the driver's age. Age is a predictive variable in almost any risk model and is accepted in tariffs. It is considered normal that younger drivers pay higher prices as they are less experienced.
4. An ongoing evolution is to have more detailed geographically segmented insurance tariffs on the postal code level or even more granularly on the neighborhood level with the help of demographic data. This could lead to higher charges for disadvantaged neighborhoods if these reveal higher claim costs, for instance due to more frequent car thefts.

These examples show that discrimination is not black and white. What we consider discrimination is determined by laws and by society. With respect to the insurance sector, gender discrimination is illegal whereas age discrimination is generally accepted.

Demand-based discrimination occurs when certain market segments are charged higher prices because they are less price sensitive. Often, this type of discrimination takes place at the renewal stage: prices are typically increased during the annual renewal of the policy. This is necessary to cope with inflation. However, price increases beyond inflation are also common, taking advantage of the fact that not every client will bother to "shop around" each year, as predicted by demand models. Note that the Financial Conduct Authority (FCA), the body regulating the English insurance market, has banned differential pricing between new business and renewals since 01/01/2022. In the EU, there is no such regulation, but individual members such as Hungary (MTPL) have taken similar steps.

The third category is *intermediary discrimination*. Note that intermediaries in the EU are usually paid a commission which is a percentage of the premium paid by the client, potentially with extra bonuses if targets are met. This can lead to incentives which are not aligned with the interests of clients. However, this is not the type of discrimination I want to discuss here (note that IDD directive 2016/97 of the EU regulates the insurance distribution). Intermediaries can sometimes directly influence the end price for the client, by giving a certain level of commercial discount (usually a percentage discount from the tariff price). Often, these discounts are granted following market circumstances, but there can be discriminatory aspects as well. Discounts can be granted based on personal relationships or certain social preferences of the intermediary, hereby discriminating other (groups of) clients. Even if intermediaries are not responsible for setting tariff prices, not granting a certain discount can also be discriminatory. Such discrimination is very hard to measure. It also forms a potential loophole for the types of discrimination mentioned above, like gender or ethnic discrimination.

Which Role does AI Play in Reinforcing/Mitigating the Discussed Types of Discrimination? In a first step, it is important to emphasize that AI is dependent on the data it is fed with. The advent of AI coincides with an evolution toward (much) more available data and the ability to include data from non-conventional sources. In turn, AI can also play a role in sourcing data (see, for instance, text mining). In my opinion, the influence of AI is often mixed up with the influence stemming from more and better data, without necessarily using the term "big data". Going back to the risk modelling stage, one should be careful not to include any discriminatory variables into the dataset. In a traditional world, the pricing actuary will make sure not to use gender in a tariff, even if it is included in the dataset or possibly in the underlying risk model. By doing so, he or she limits the risk of direct discrimination.

The picture looks different when talking about indirect discrimination: AI might be able to spot certain effects that the pricing actuary does not, especially when these effects deal with interactions between two or more variables. For example, indirect ethnic discrimination could occur by including correlated variables such as income, level of education, employment, or others. When using AI techniques, there is a higher risk of indirectly discriminatory variables "sneaking" into the model through interactions. Therefore, the careful monitoring of the variables in the model is vital. It is also important whether the AI model is used as a "final model" or rather as a "helper" for more traditional models. However, the core challenge remains unchanged: there is a clear necessity on determining what is discrimination and what is not.

Let us now ask the same question for demand-based discrimination. Whereas risk is rather stable over time, demand is much more dynamic: if our main competitor decides to drop prices by 10% tomorrow, the demand model we just created already needs an update. For models which are refreshed more frequently, AI offers large productivity gains compared to traditional techniques. Entire processes can be automated and model actualizations can take place instantly. Consequently, these models are typically less deeply analyzed by the pricing actuary. Hence, the risk of indirect discrimination mentioned above is more present. For our third risk, intermediary discrimination, AI could have an indirect positive impact. The reason lies not in the techniques themselves but can be attributed to the fact that, when tariffs become more precise and sophisticated, typically, the discount competences for the intermediaries are reduced. Indeed, investing a lot of time and money to get a tariff up to 1 EUR "optimized", while allowing intermediaries to grant 10–20% discounts would seem counterintuitive. This trend toward fewer discounts is also influenced by the shift to selling insurance online. However, large differences between products and countries continue to exist here.

As a general conclusion, AI may exacerbate certain already inherent forms of discrimination, but whether real discrimination takes place largely remains subject to human decisions. It would be wrong to focus on AI as the main cause of discrimination as discrimination can also take place in a very traditional setting. The discussion is surely not yet in its final stages, considering that EIOPA, the European insurance authority, has picked up on the topic of "differential pricing practices" as well.

6 Insights from Philosophy and Ethics

Suzana Alpsancar

Digital ethics has three main objectives: a diagnostic analysis, a practical evaluation, and a theoretical justification. The practical aim is to deliver a proactive and retrospective evaluation of the use of technology in their respective contexts (Jacobs et al. 2021). The theoretical aim is to provide and justify arguments, criteria, or principles that provide an orientation for the practical evaluations (Sollie 2007). Given the high context-sensitivity of digital ethics, we need to start by thoroughly analyzing the case at hand (diagnostic analysis) for each consideration: To investigate which specific difference the implementation of algorithmic decision-making (ADM) makes, the respective socio-technical contexts have to be analyzed thoroughly. Which particular challenges regarding discrimination do we face because we are using ADM instead of other means?

6.1 What are We Dealing With?

Only 16 h after its release on Twitter on March 23 in 2016, Microsoft Corporation pulled back its chatbot Tay.ai, which had quickly gained more than 500,000 followers and posted over 100,000 tweets. Many were inflammatory or even derogatory attacks against Jews, People of Color, or women (Reese 2016; Vincent 2016). Stating that some users had exploited Tay's technical vulnerabilities, which they did not foresee but took responsibility for, Microsoft declared Tay to be a social as well as a technical experiment necessary to advance AI: "To do AI right, one needs to iterate with many people and often in public forums" (Lee 2016). This example shows that it is not always easy to determine whether or not a digital service is market-ready.

The question of reliability is complicated with regard to adaptive systems meant to further optimize themselves once out in the wild. Some ADM have incorrectly influenced grave decisions such as the probability of death of a patient with pneumonia (Caruana et al. 2015; Cabitza et al. 2017) or have been subject to adversarial attacks (Gilpin et al. 2018), while others have been easy to trick and have exhibited Clever-Hans effects (Kraus and Ganschow 2022), domain shifts, or overfitting (Cremers et al. 2019; Ribeiro et al. 2016). How can the public then be sure that ADM systems have been tested and validated sufficiently? Do we need certifications (Krafft et al. 2022; Möslein and Zicari 2021) in general or just for those classified as high-risk systems according to the EU AI Act (European Commission 2021)?

Beyond this peculiar product status of software, most of the systems contributing to today's success of ADM are opaque, meaning that, for a variety of reasons, it is not (immediately) obvious how they work or why they exhibit a particular behavior or performance (Burrell 2016; Creel 2020; Resch and Kaminski 2019; Sullivan 2020). While opacity due to corporate secrecy can, in principle, be regulated, opacity due to intrinsic technical features can become problematic in terms of accountability. Usually, to hold someone responsible for some decision implies that this someone had a meaningful understanding of how this decision was made. Accordingly, some sort of transparency or explainability is often seen as mandatory to enable accountability (Floridi et al. 2018) and to make sure that those potentially affected can somehow determine if they have or have not been subject to unfair decision-making (Dotson 2014; Benjamin 2019). Moreover, opacity itself might be seen as discriminating, as software is not mutually opaque to everyone (Zednik 2021). Instead, this varies

according to the degree of illiteracy and information and power asymmetry (Burrell 2016; Lepri et al. 2018). A potential political issue in the future may be who has access to (good) digital services – e.g., in administration, health care, or education – and who does not, paving the way for a "digital divide" (Boyd and Crawford 2012), either because people are subjected to unfair ADM unevenly or because they benefit from the systems unevenly.

6.2 Transformations of Socio-Technical Constellations

If we want to understand how people might be affected by using ADM we must consider the different social positions, roles, and constellations in which the ADM are being implemented and how these might be transformed. For instance, using ADM for recommending medical treatment directly mediates (Verbeek 2005) the doctor–patient relationship but can also alter the relationship between members of a team of physicians in a clinic, as well as their relationship to the patient's family members, to other patients, or to the medical care system as such (e.g., If there is reliable ADM for detecting cancer, should all insured people have a right to be diagnosed by these machines?). In consequence, those whose workplaces adopt ADM have to readjust their role as professionals and find themselves in the new responsibility of deciding when, and when not, to rely on the machine (de Visser et al. 2020; Schaffer et al. 2019).

Given that roles and constellations vary throughout different workplaces, we need to thoroughly account for all particular perspectives of each case. There are deviating categorizations of stakeholders in the literature (Arrieta et al. 2020; Preece et al. 2018; Zednik 2021; Dhanorkar et al. 2021). Here, the most typical groups are displayed: Developers (such as engineers, data scientists, product owners, companies, managers, executive board members, and alike), distributors (such as retailers and dealers), operators (such as domain experts or users of ADM), clients (often those affected by the model such as patients or customers), and, finally, regulators (such as governmental agencies, NGOs, or civil associations).

While there is a growing consensus that engineers and developers should try to include different stakeholders' views (e.g., by participatory design, see Dignum 2019; Neuhauser and Kreps 2011), it is equally important to add the normative position of the public. The public occupies an ideal position that calls for a specific form of reflection: the task to check for intersubjective justifiability. We may think of the public in terms of citizens of a particular state or society, the people of a cultural community, or even in the sense of humankind. While stakeholders' interests, needs, or demands can be investigated empirically as, for instance, proposed in Value-Sensitive-Design approaches (Van de Poel 2020), the normative position of the public relates to the idea of changing perspectives, of being impartial, or of judging from a universal point of view. This normative idea should be used as a critical tool that allows us to assess the goodness of certain normative claims. For instance, Rawls (1999), who conceptualizes justice as fairness, famously evoked the so-called "veil-of-ignorance" – a thought experiment for evaluating the fairness of social institutions. Put simply, he calls to ask ourselves: If you did not know where you were standing in a society (or in the world; e.g., in terms of place of birth, race, sex, gender, age, profession, capital, or other criteria), would you hold claim x to be fair?

6.3 Large-Scale and Long-Time Effects

There are two major concerns regarding large-scale and long-time effects of ADM. In light of a market with many different companies and state agencies on the demand side and few players on the supply side, Creel and Hellman (2022) argue that standardized ADM replace or influence thousands of unique human deciders who, before, based their decisions on multiple and diverse criteria. Using standardized ADM means a homogenization of how decisions are made: First by formalizing the process completely to be processable by algorithms and then by using the same model within or even beyond a societal sector. If, for instance, such a model were to discriminate against People of Color, this discrimination would not only take place locally, e.g., in the hiring process of a particular company, but would by definition reject the same group of people everywhere that that software is in use. In the extreme case, this group would then be denied any chance of being hired.

The second concern is that discriminatory outcomes can be "self-reinforcing", meaning that those who have been disadvantaged in the past will also be disadvantaged in the present and even more severely in the future (O'Neil 2016; Benjamin 2019). For example, assuming that having a high school degree is favorable for decisions regarding one's creditworthiness and that it is known that disabilities reduce the chances of achieving higher education, a group that is disadvantaged in one sector of society may also have lower chances in another sector of society. If you are not able to receive a credit, you may also not be able to rent or buy a house in a good neighborhood, which in turn may lower your chances of getting hired in a good company as well as your children's chance of being admitted to the school of your choosing. In the long run, this can lead to chain effects of discrimination that run counter to the principle of equal opportunity (Lepri et al. 2018) – one of the prime promises of modern, liberal societies and current social politics.

In conclusion, we have seen that the issue of fairness, bias, and discrimination interacts with other ethical and societal concerns such as opacity, power, autonomy, and accountability but also with questions of privacy and cybersecurity, which could not be further elaborated on here. Accordingly, the ethical discussion should not be limited to designing for fairness. Further, fairness should not be conceptualized as a property or feature of a technical artifact alone but rather of a whole sociotechnical system (Selbst et al. 2019; Suchman and Suchman 2006). Respectively, we should acknowledge that what counts as fair or unfair is not only highly context-sensitive but also always contestable – for good reasons: If we follow the democratic idea that we are not all the same but want to live in a just society (e.g., equal opportunities for all), then we will always have to deal with biased choices and institutions. The best one can do is to explicate all relevant decisions and open them up for debate, while being aware that what seems to be the best possible fair solution today may not appear to be so in the (near) future. Consequently, we should ensure the possibility of reassessing sociotechnical systems in the future, thereby avoiding lock-in effects (D21 2020).

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Allan TRS (1998) Procedural fairness and the duty of respect. Oxf J Leg Stud 18:497–515

Alufaisan Y, Marusich LR, Bakdash JZ, Zhou Y, Kantarcioglu M (2021) Does explainable artificial intelligence improve human decision-making? Proc AAAI Conf Artif Intel 35:6618–6626. https://doi.org/10.1609/AAAI.V35I8.16819

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. In: Ethics of data and analytics. Auerbach, pp 254–264

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115

Barocas S, Hardt M, Narayanan A (2018) Fairness and machine learning. https://fairmlbook.org/. Accessed 01 Jan 2023

Benjamin R (2019) Race after technology. Abolitionist tools for the new Jim Code. Polity Press, Cambridge

Bomhard D, Merkle M (2021) Regulation of artificial intelligence. J Eur Consum Mark Law 21:257–262

Boyd D, Crawford K (2012) Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf Commun Soc 15:662–679. https://doi.org/10.1080/1369118X.2012.678878

Burrell J (2016) How the machine thinks: understanding opacity in machine learning algorithms. Big Data Soc 3:1–12. https://doi.org/10.1177/2053951715622512

Cabitza F, Rasoini R, Gensini GF (2017) Unintended consequences of machine learning in medicine. JAMA 318:517–518. https://doi.org/10.1001/jama.2017.7797

Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N (2015) Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1721–1730. https://doi.org/10.1145/2783258.2788613

Castelnovo A, Crupi R, Greco G, Regoli D, Penco IG, Cosentini AC (2022) A clarification of the nuances in the fairness metrics landscape. Sci Rep 12:1–21. https://doi.org/10.1038/s41598-022-07939-1

Caton S, Haas C (2020) Fairness in machine learning: a survey. ArXiv preprint. http://arxiv.org/abs/2010.04053. Accessed 01 Jan 2023

Chen IY, Johansson FD, Sontag D (2018) Why is my classifier discriminatory? In: Proceedings of the 32nd international conference on neural information processing systems, pp 3543–3554

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

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp 797–806

Creel K (2020) Transparency in complex computational systems. Philos Sci 87:568–589. https://doi.org/10.1086/709729

Creel K, Hellman D (2022) The algorithmic leviathan: arbitrariness, fairness, and opportunity in algorithmic decision-making systems. Can J Philos 52:26–43. https://doi.org/10.1017/can.2022.3

Cremers, AB, Englander A, Gabriel M, Hecker D, Mock M, Poretschkin M, Rosenzweig J, Rostalski F, Sicking J, Volmer J, Voosholz J, Voss A, Wrobel S (2019) Trustworthy use of artificial intelligence. Priorities from a philosophical, ethical, legal and technological viewpoint as a basis for certification of artificial intelligence. In: Whitepaper. Fraunhofer Institute for Intelligent Analysis and Information Systems

D21 (2020) Denkimpuls Digitale Ethik: bias in algorithmic systems explanations, examples and arguments. Tech. rep. initiative D 21. Working group monitoring of algorithms. https://initiative d21.de/publikationen/denkimpulse-zur-digitalen-ethik. Accessed 10 Oct 2022

Dastile X, Celik T, Potsane M (2020) Statistical and machine learning models in credit scoring: a systematic literature survey. Appl Soft Comput. https://doi.org/10.1016/j.asoc.2020.106263

De Visser EJ, Peeters MMM, Jung MF, Kohn S, Shaw TH, Pak R, Neerincx MA (2020) Towards a theory of longitudinal trust calibration in human–robot teams. Int J Soc Robot 12:459–478. https://doi.org/10.1007/s12369-019-00596-x

Dhanorkar S, Wolf CT, Qian K, Xu A, Popa L, Li Y (2021) Who needs to know what, when? Broadening the explainable AI (XAI) design space by looking at explanations across the AI lifecycle. In: Proceedings of the designing interactive systems conference, pp 1591–1602. https://doi.org/10.1145/3461778.3462131.3

Dodge J, Liao QV, Zhang Y, Bellamy RKE, Dugan C (2019) Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th international conference on intelligent user interfaces 11. https://doi.org/10.1145/3301275

Dignum V (2019) Responsible artificial intelligence: artificial intelligence: foundations, theory, and algorithms. Springer, Cham. https://doi.org/10.1007/978-3-030-30371-6_1

Dolata M, Feuerriegel S, Schwabe G (2021) A sociotechnical view of algorithmic fairness. Inf Syst J 32(4):754–818. https://doi.org/10.1111/ISJ.12370

Dotson K (2014) Conceptualizing epistemic oppression. Soc Epistemol 28:115–138. https://doi.org/10.1080/02691728.2013.782585

Ebers M, Hoch V, Rosenkranz F, Ruschemeier H, Steinrötter B (2021) Regulation of artificial intelligence. Multidiscipl Sci J 21:589–601

Ellis E, Watson P (2012) EU anti-discrimination law, 2nd edn. Oxford University Press, Oxford

European Commission (2021) Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM (2021) 206. https://ec.europa.eu/transparency/documents-register/detail?ref=COM(2021)206&lang=en. Accessed 22 Nov 2022

Feuerriegel S, Dolata M, Schwabe G (2020) Fair AI—challenges and opportunities. Bus Inf Syst Eng 62:379–384. https://doi.org/10.1007/s12599-020-00650-3

Flores AW, Bechtel K, Lowenkamp CT (2016) False positives, false negatives, and false analyses: a rejoinder to machine bias: there's software used across the country to predict future criminals and it's biased against blacks. Fed Probat 80:38

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach 28:689–707. https://doi.org/10.1007/s11023-018-9482-5

Fu R, Aseri M, Singh PV, Srinivasan K (2022) "Un" fair machine learning algorithms. Manag Sci 68:4173–4195. https://doi.org/10.1287/mnsc.2021.4065

Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: an approach to evaluating interpretability of machine learning. In: Proceedings of the 2018 IEEE 5th international conference on data science and advanced analytics

Haas C (2019) The price of fairness—a framework to explore trade-offs in algorithmic fairness. In: Proceedings of the international conference on information systems

Hleg AI (2018) Ethics guidelines for trustworthy AI. https://digitalstrategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed 22 Nov 2022

Hutchinson B, Mitchell M (2019) 50 years of test (un)fairness: lessons for machine learning. In: Proceedings of the conference on fairness, accountability, and transparency, pp 49–58. https://doi.org/10.1145/3287560.3287600

Ince ST (2021) European Union law and mitigation of artificial intelligence-related discrimination risks in the private sector: with special focus on the proposed Artificial Intelligence Act. Istanbul University Press. https://dergipark.org.tr/en/download/article-file/1912427. Accessed 22 Nov 2022

Jacobs M, Kurtz C, Simon J, Böhmann T (2021) Value sensitive design and power in socio-technical ecosystems. Internet Policy Rev 10:1–26. https://doi.org/10.14763/2021.3.1580

Kaplow L, Shavell S (2002) Fairness and welfare. Harvard University Press, Cambridge

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2017) Human decisions and machine predictions. Q J Econ 133(1):237–293. https://doi.org/10.3386/w23180

Konow J (2009) Is fairness in the eye of the beholder? An impartial spectator analysis of justice. Soc Choice Welf 33:101–127

Kordzadeh N, Ghasemaghaei M (2022) Algorithmic bias: review, synthesis, and future research directions. Eur J Inf Syst 31:388–409

Krafft TD, Zweig KA, König PD (2022) How to regulate algorithmic decision-making: a framework of regulatory requirements for different applications. Regul Gov 16:119–136. https://doi.org/10.1111/rego.12369

Kraus T, Ganschow L (2022) Anwendungen und Lösungsansätze erklärbarer Künstlicher Intelligenz. In: Hartmann EA (ed) Digitalisierung souverän gestalten II. Springer, Heidelberg, pp 38–50

Lambrecht A, Tucker C (2019) Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Manag Sci 65:2966–2981

Lee P (2016) Learning from Tay's introduction. In: Official microsoft blog. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/. Accessed 10 Oct 2022

Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2018) Fair, transparent, and accountable algorithmic decision-making processes. Philos Technol. https://doi.org/10.1007/s13347-017-0279-x

Mahler T (2022) Between risk management and proportionality: the risk-based approach in the EU's Artificial Intelligence Act. Nord Yearb Law Inform 2022:247–271

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54:1–35

Menon AK, Williamson RC (2018) The cost of fairness in binary classification. In: Proceedings of machine learning research, vol 81, pp 107–118. https://proceedings.mlr.press/v81/menon18a.html

Mitchell S, Potash E, Barocas S, D'Amour A, Lum K (2021) Algorithmic fairness: choices, assumptions, and definitions. Ann Rev Stat Appl 8:141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Möslein F (2022) AI and corporate law. In: DiMatteo LA, Poncibò C, Cannarsa M (eds) The Cambridge handbook of artificial intelligence. Cambridge University Press, Cambridge, pp 74–86

Möslein F, Horn M (2021) Emerging rules on artificial intelligence: Trojan horses of ethics in the realm of law? In: DiMatteo LA, Janssen A, Ortolani P, de Elizalde F, Cannarsa M, Durovic M (eds) The Cambridge handbook of lawyering in the digital age. Cambridge University Press, Cambridge, pp 77–95

Möslein F, Zicari RV (2021) Certifying artificial intelligence systems. In: Vogl R (ed) Research handbook on big data law. Elgar, Cheltenham. https://doi.org/10.4337/9781788972826.00024

Neuhauser L, Kreps GL (2011) Participatory design and artificial intelligence: strategies to improve health communication for diverse audiences. In: AAAI spring symposium: AI and health communication

O'Neil C (2016) Weapons of math destruction. How big data increases inequality and threatens democracy. Crown, New York. https://doi.org/10.1177/0256090919853933

Peterson M (2017) The ethics of technology: a geometric analysis of five moral principles. Oxford University Press, Oxford

Preece A, Harborne D, Braines D, Tomsett R, Chakraborty S (2018) Stakeholders in explainable AI. arXiv preprint. http://arxiv.org/abs/1810.00184. Accessed 12 Dec 2022

Raghavan M, Barocas S, Kleinberg J, Levy K, Levy K (2020) Mitigating bias in algorithmic hiring: evaluating claims and practices. In: Proceedings of the 2020 conference on fairness,

accountability, and transparency. https://doi.org/10.1145/3351095.3372828
Rawls J (1958) Justice as fairness. Philos Rev 67:164–194
Rawls J (1999) A theory of justice. Belknap Press, Cambridge
Reese H (2016) Why Microsoft's 'Tay' AI bot went wrong. In: TechRepublic. https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/. Accessed 10 Oct 2022
Resch M, Kaminski A (2019) The epistemic importance of technology in computer simulation and machine learning. Mind Mach 29:9–17. https://doi.org/10.1007/s11023-019-09496-5
Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco, pp 1135–1144. https://doi.org/10.1145/2939672.2939778
Sadurski W (2011) Commutative, distributive and procedural justice. The many concepts of social justice in European private law. Elgar, Cheltenham, pp 90–104
Schaffer J, O'Donovan J, Michaelis J, Raglin A, Höllerer T (2019) I Can do better than your AI: expertise and explanations. In: Proceedings of the 24th international conference on intelligent user interfaces, pp 240–251. https://doi.org/10.1145/3301275.3302308
Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: Proceedings of the conference on fairness, accountability, and transparency, pp 59–68. https://doi.org/10.1145/3287560.3287598
Shulner-Tal A, 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 Inf Technol 24:1. https://doi.org/10.1007/S10676-022-09623-4
Sollie P (2007) Ethics, technology development and uncertainty: an outline for any future ethics of technology. J Inf Commun Ethics Soc 5:293–306. https://doi.org/10.1108/14779960710846155
Suchman L, Suchman LA (2006) Human-machine reconfigurations plans and situated actions. Cambridge University Press. https://doi.org/10.1017/CBO9780511808418
Sullivan E (2020) Understanding from machine learning models. Br J Philos Sci 73(1):109–133. https://doi.org/10.1093/bjps/axz035
Sweeney L (2013) Discrimination in online ad delivery. Commun ACM 56:44–54
Townsend B (2021) Decoding the proposed European Union Artificial Intelligence Act. Am Soc Int Law 25:20
Van de Poel I (2020) Embedding values in artificial intelligence (AI) systems. Mind Mach 30:385–409. https://doi.org/10.1007/s11023-020-09537-4
Veale M, Borgesius FZ (2021) Demystifying the draft EU Artificial Intelligence Act. Comput Law Rev Int 22:97–112
Verbeek P-P (2005) What things do: philosophical reflections on technology, agency, and design. Pennsylvania State University Press, University Park
Verma S, Rubin J (2018) Fairness definitions explained. In: Proceedings of the 2018 IEEE/ACM international workshop on software fairness, pp 1–7
Vincent J (2016) Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. In: The Verge. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist. Accessed 10 Oct 2022
Willett C (2007) Fairness in consumer contracts. Ashgate, Aldershot
Yona G (2017) A gentle introduction to the discussion on algorithmic fairness. Towards Data Sci 5. https://towardsdatascience.com/a-gentleintroduction-to-the-discussion-on-algorithmic-fairness-740bbb469b6. Accessed 24 Jan 2023
Zednik C (2021) Solving the black box problem: a normative framework for explainable artificial intelligence. Philos Technol 34:265–288. https://doi.org/10.1007/s13347-019-00382-7.5