AI & SOCIETY (2022) 37:143–154
https://doi.org/10.1007/s00146-021-01192-2

ORIGINAL ARTICLE

# Discrimination in the age of artificial intelligence

**Bert Heinrichs**[1,2]

Received: 28 September 2020 / Accepted: 16 March 2021 / Published online: 4 April 2021
© The Author(s) 2021

**Abstract**
In this paper, I examine whether the use of artificial intelligence (AI) and automated decision-making (ADM) aggravates issues of discrimination as has been argued by several authors. For this purpose, I first take up the lively philosophical debate on discrimination and present my own definition of the concept. Equipped with this account, I subsequently review some of the recent literature on the use AI/ADM and discrimination. I explain how my account of discrimination helps to understand that the general claim in view of the aggravation of discrimination is unwarranted. Finally, I argue that the use of AI/ADM can, in fact, increase issues of discrimination, but in a different way than most critics assume: it is due to its epistemic opacity that AI/ADM threatens to undermine our moral deliberation which is essential for reaching a common understanding of what should count as discrimination. As a consequence, it turns out that algorithms may actually help to detect hidden forms of discrimination.

**Keywords** Discrimination · Artificial intelligence · Deep learning · Transparency

## 1 Introduction

The argumentation presented in this paper proceeds in three steps: in part one (Sect. 2–5), I will take up the rich and sophisticated philosophical discussion on discrimination of recent years and take sides with those who criticize the presumption that the concept carries significant normative power. In fact, I will push this criticism a step further and argue that the notion of discrimination itself is inappropriate for ethical justification. I will argue that it can only serve to flag actions which we consider morally wrongful *for other reasons*. This does not mean that the notion of discrimination is entirely useless. It rather means that one must pay attention to whether the notion is applied properly. The correct use of the notion of discrimination is "Action A is a case of discrimination, *because of* ω [where ω refers to an established weighting of relevant ethical concerns in a given context]" and not "Action A is wrong, *because* it is a case of discrimination".

Equipped with this line of argument I will, in part two (Sect. 6 and 7), review some of the recent literature on the use of artificial intelligence and automated decision-making (henceforth AI/ADM) in various areas of life. Several authors have claimed that the use of AI/ADM aggravates issues of discrimination. I will maintain that the notion of discrimination is frequently used in an unqualified manner as if it were, by itself, a source of ethical justification. However, given the analysis of part one, a reasonable weighting of relevant ethical concerns needs to be provided to show that the use of AI/ADM in a particular area is ethically objectionable.

In the final part (Sect. 8 and 9), I will claim that the use of AI/ADM can, in fact, aggravate issues of discrimination, but in a different way than most critics assume. With the implementation of AI/ADM, it becomes much more difficult to determine whether certain instances are at odds with an established weighting of relevant ethical concerns. It is due to its epistemic opacity that AI/ADM threatens to undermine exactly the kind of moral deliberation that is essential for reaching an understanding in view of what should count as discrimination. However, algorithms may also help to detect hidden forms of discrimination.

✉ Bert Heinrichs
b.heinrichs@fz-juelich.de

[1] Institute of Neuroscience and Medicine, Ethics in the Neurosciences (INM-8), Forschungszentrum Jülich, 52425 Jülich, Germany

[2] Institute of Science and Ethics (IWE), Rheinische Friedrich-Wilhelms-Universität Bonn, Bonner Talweg 57, 53113 Bonn, Germany

## 2 The philosophical debate about the concept of discrimination

A couple of years ago, a vigorous debate about the concept of discrimination commenced in philosophy. Although commonly used in everyday talk and prevalent in many national legislations and supra-national codes—not least in article 7 of the *Universal Declaration of Human Rights*—, providing a conclusive analysis of the concept of discrimination proved to be difficult. Defining discrimination in a way that covers all or at least most of the established uses turned out to be especially hard. Approaches which simply take discrimination to denote unjust differential treatment are, in contrast, uninformative. Moreover, controversies arose regarding the question why exactly discrimination is morally wrong.[1] Notwithstanding lasting disagreements, the debate has undoubtedly brought about considerable clarification. In particular, it has led to a number of conceptual differentiations which are now well established. These include, above all, the distinctions between direct and indirect discrimination as well as between individual, organizational and institutional discrimination (cf. Altman 2020, Sec. 2). However, for the argumentation in part one (Sect. 2–5) of this paper these distinctions are less important. The following discussion in part two (Sect. 6 and 7) and three (Sect. 8 and 9) is primarily targeted at discrimination by organizations, but its scope is not strictly limited to this form of discrimination.

Regarding the *problem of definition*, Kasper Lippert-Rasmussen has formulated what may now be called the standard view:

> "X discriminates against (in favour of) Y in dimension W iif: (i) X treats Y differently from Z (or from how X would treat Z, were X to treat Z in some way) in dimension W; (ii) the differential treatment is (or is believed by X to be) disadvantageous (advantageous) to Y; and (iii) the differential treatment is suitably explained by Y's and Z's being (or believed by X to be) (members of) different, socially salient groups." (Lippert-Rasmussen 2006, 168)

Note that this definition is non-moralized, i.e., it covers both wrongful and non-wrongful forms of discrimination. In particular, for Lippert-Rasmussen as well as for many other scholars, discrimination is only contingently bad (Lippert-Rasmussen 2006, 174). In contrast, I will adopt a moralized definition below according to which discrimination is necessarily bad (for the distinction between moralized/non-moralized understandings of discrimination see Lippert-Rasmussen 2013, 24–26; Thomsen 2018, 26–27).

While it was obvious before that discrimination is closely linked to group membership, it was not at all clear what kind of group membership qualifies for discriminatory practices. With his approach, Lippert-Rasmussen suggested that "social saliency" is essential in this regard. A typical legal provision on discrimination like the one in article 26 of the *International Covenant on Civil and Political Rights* includes a list of personal traits:

> "All persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status." (ICCPR, Art. 26)

According to Lippert-Rasmussen, the common feature of the traits mentioned—race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status—lies in the fact that the groups defined by these traits are "socially salient". In view of this characterization, he explicates:

> "A group is socially salient if perceived membership of it is important to the structure of social interactions across a wide range of social contexts." (Lippert-Rasmussen 2006, 169)

On this account, it is a case of discrimination (if it is) if an employer favors applicants, for example, on the basis of their race, say whites, and rejects others. It might also be wrong if an employer favors applicants with a certain eye color, say green, and rejects those with blue and brown eyes. However, according to Lippert-Rasmussen, such a behavior would not qualify as discrimination because the group of green-eyed people is not socially salient in the sense specified while the group of whites (and blacks respectively) is. Some consider this as an unreasonable limitation of the concept of discrimination (Thomsen 2013).

The main advantage of Lippert-Rasmussen's approach is that it allows to dismiss those traits which are commonly not associated with discrimination. Lippert-Rasmussen mentions "non-family-members, unqualified applicants, or the undeserving" (2006, 169) as examples of such traits. If an employer rejects candidates on the basis of lacking qualifications, we would hardly call this discrimination. Following Lippert-Rasmussen, the reason for this is exactly that this

[1] Important contributions to this debate include Alexander (1992), Halldenius (2005), Lippert-Rasmussen (2006), Arneson (2006), Heinrichs (2007), Hellmann (2008), Moreau (2010), Segall (2012), Thomsen (2013), Lippert-Rasmussen (2013) and Cook (2015). In addition, some handbook articles have been influential, especially Ezorsky (1992), Nickel (1998) and Wasserman (1998). Altman (2020) provides an excellent systematic overview of this debate. Lippert-Rasmussen (2018) has recently edited a comprehensive anthology.

trait does not define a socially salient group. Social saliency is, so to speak, the recurrent theme that connects those traits which define group memberships which are relevant regarding discrimination.

Apparently, the concept of discrimination is, on this approach, a relative one: Social saliency may change over time and is dependent on cultural circumstances. This is consistent with the observation that a certain form of differential treatment may count as discrimination in one context and may be acceptable in another. It might, for example, be that in a particular society green-eyed and red-haired people have long suffered from disadvantageous treatment. Arguably, in such a society, the above example of an employer who rejects an applicant because of his green eyes should count as discrimination just as disadvantageous treatment because of race or gender does count as discrimination in most modern societies.

Regarding the *question of the moral wrongness* of discrimination, Lippert-Rasmussen has advocated a harm-based account as compared to a disrespect-based account (Lippert-Rasmussen 2006, 174–184). The latter has, among others, been supported by Deborah Hellman. She maintains:

> "Discrimination is wrong when it demeans. To demean is to treat another as less worthy." (Hellman 2008, 33)

While on first view, Hellman's approach seems to capture the particular kind of wrongness of discriminatory acts, Lippert-Rasmussen has argued against it on grounds that a discriminator's false views about moral status are inappropriate to explain the moral badness at stake (Lippert-Rasmussen 2006, 182–184). Rather, according to him, it is the fact of inflicted harm that accounts for the moral wrongness of discrimination. In the course of the debate, other accounts have been developed. Moreau (2010), for example, has maintained that discrimination is wrong because it interferes with deliberative freedoms which are equally important for everyone. Segall (2012), in turn, has argued that discrimination is bad because it undermines the equality of opportunity. Despite the differences, all the authors agree that (wrongful) discrimination is a distinctive form of moral wrongdoing that targets people because they belong to some group while other people which do not belong to this group are spared.

## 3 A critique of the social saliency account

Undeniably, Lippert-Rasmussen and others have pushed the debate about discrimination significantly further. However, some critics have claimed that his definition is both, too inclusive and too exclusive at the same time. On their view, the group criterion in general is not suitable to separate traits which are relevant in the context of discrimination (given a widely shared understanding) from those which are not. Most notably, Frej Thomsen has argued against a definition of discrimination including the group criterion. According to Thomsen, there are two strategies for justifying the group criterion (Thomsen 2013, 13ff. and 23ff.): either one tries to establish that there are "inherently relevant groups" or one focuses on "contextually relevant groups" (the latter being more or less Lippert-Rasmussen's strategy). Yet, according to Thomsen, both strategies lead to a dead end. A natural way for further elaborating the first strategy is to spell out "inherent relevance" in terms of "responsibility for possessing a trait" (Thomsen 2013, 14). As Thomsen shows, a further differentiation is required to evaluate this account. Responsibility for having a trait can mean "past-responsibility", "future-responsibility", or "immutability" (Thomsen 2013, 17). However, none of these accounts matches well with what we ordinarily consider as typical traits in the context of discrimination. Take immutability, for example: while race, color, national or social origin, and birth are immutable, language, religion, political or other opinion, property, and other status are not (and sex is somewhat undecided). Nevertheless, we expect both groups to be on the list of traits which are relevant for discrimination—at least in some contexts. Moreover, there are other traits which are immutable—e.g., shoe size—which are typically not on the list of relevant traits. Similar counterexamples can be found for past-responsibility and future-responsibility. Consequently, one must admit that all three categories are both, too inclusive and too exclusive at the same time. Hence, all these accounts fail to allow for specifying a set of traits that characterize group membership in a way that meets the received extension of discrimination.

The second strategy might appear more promising since it focusses from the beginning "on a relatively well-defined set of traits, which are constitutive of groups that stand out because of their socially and historically specific group identity." (Thomsen 2013, 23) However, according to Thomsen, this strategy introduces a new problem, because on this account idiosyncratic uses of certain traits for treating people differently are excluded from the concept of discrimination. While such acts may still be morally objectionable, they do not qualify as cases of discrimination. In fact, Lippert-Rasmussen explicitly acknowledges this and argues that it is in line with a common understanding of discrimination (Lippert-Rasmussen 2006, 169). Therefore, he could argue that Thomsen's criticism falls short.

Yet, there is another problem with the second strategy, namely that referring to membership in a socially salient group does not constitute a case of discrimination under all circumstances and in all contexts. Imagine that a manufacturer of bikini wear is seeking models for presenting the new collection (Heinrichs 2007, 106). Rejecting male applicants would hardly count as an instance of discrimination,

although gender is typically included in the list of relevant traits. Arguably, we would consider the manufacturer's interests to have an appropriate promotion of his collection well founded. One could continue to argue that there is a "factual link" between the activity (promotion of bikini wear) and the trait in question (being female) (Heinrichs 2007, 107). This would be in line with most anti-discrimination regulations which typically include exceptions. The EU Employment Equality Directive (2000/78/EC), for example, holds in Art. 4 (1) that a difference of treatment which is based on discrimination-relevant characteristics such as sex "shall not constitute discrimination where, by reason of the nature of the particular occupational activities concerned or of the context in which they are carried out, such a characteristic constitutes *a genuine and determining occupational requirement*, provided that the objective is legitimate and the requirement is proportionate." (italics added) (cf. European Union Agency for Fundamental Rights 2011, sec. 2.6.4) On closer inspection, however, it becomes clear that the link between activity or context on the one hand and personal trait on the other is never purely "factual" but always includes normative assumptions about appropriateness and reasonableness. We may, for example, find it appropriate that an employer expends considerable effort on creating a work environment that is suitable for a disabled employee. The claim that the employee is unable to pursue the activity without costly modifications would hardly count as a sufficient reason for rejecting him or her. Overall, this proves that social saliency alone is neither necessary nor sufficient for defining what discrimination is.

## 4 A revised definition of discrimination

Given the central role that the notion of socially salient groups plays in Lippert-Rasmussen's account of discrimination, its rejection has a far-reaching implication. Apparently, it is not a set of traits alone which defines what discrimination is. However, a revised definition of discrimination can preserve the insights of Lippert-Rasmussen's account and, at the same time, get rid of its shortcomings. What is needed are two modifications: (1) deleting the specification "socially salient" in Lippert-Rasmussen's third qualification and (2) adding a fourth, namely that the treatment is not in accordance with an established *weighting* of various relevant ethical concerns.

The notion "relevant ethical concerns" refers to accepted prima facie principles such as respect for the self-determination of others, do not harm, and justice, but may also include specific features of the area of life or the concrete situation in question. The notion "established weighting" indicates that it is not merely a set of ethical concerns that the concept of discrimination incorporates, but that it is rather the result of a balancing process that fuels it. Viewed in this light, it becomes clear that the concept of discrimination has *by itself* no normative power. It is only a shortcut for a normative weighting (or, more precisely, for the disregard of such a weighting) which has been reached before. Classifying something as a case of discrimination is only possible *after* such a weighting has been established, not as an argument in the course of a debate on the use or non-use of certain traits. It is important to note that once a weighting has been established the concept of discrimination cannot be balanced against other ethical principles, because to call something a case of discrimination is already an *all things considered*-statement. It is for this reason that the principle of non-discrimination sometimes has the appearance of an especially important prima facie principle. In fact, it is not a prima facie principle at all. As a side note, it may be observed that on this account of discrimination, the debate on the specific wrong-making feature is likely to be undecidable. The nature of the moral wrongness of discrimination will, at least partly, depend upon which ethical principle gained dominance in the weighting process. To discriminate against someone may, therefore, sometimes be wrong because of the harm done and sometimes because of rights not granted or the opportunities not provided.

The revised definition of discrimination reads as follows:

> "X discriminates against (in favour of) Y in dimension W iif: (i) X treats Y differently from Z (or from how X would treat Z, were X to treat Z in some way) in dimension W; (ii) the differential treatment is (or is believed by X to be) disadvantageous (advantageous) to Y; (iii) the differential treatment is suitably explained by Y's and Z's being (or believed by X to be) (members of) different groups; and (iv) the treatment is not in accordance with an established weighting ω of relevant ethical concerns."

This modified definition does not render the notion of discrimination useless. One only must pay attention to whether the notion is applied properly. The correct use of the notion of discrimination is not "Action A is wrong, *because* it is a case of discrimination", but rather "Action A is a case of discrimination, *because of* ω [where ω refers to an established weighting of relevant ethical concerns in a given context]". In many cases, ω will incorporate principles such as respect for self-determination, do not harm, and justice, but also more concrete ideas of fairness (including experiences of unfairness in the past).

Note, again, that this understanding of discrimination deviates from the understanding of Lippert-Rasmussen and others in that it takes discrimination as a *moralized* concept. As mentioned above, for Lippert-Rasmussen discrimination is only "contingently bad". He observes: "An instance of discrimination is pro tanto bad, when it is, because it makes the

discriminatees worse off." (2006, 174) This means, in turn, if an act of discrimination does not make the discriminatees worse off it is still an act of discrimination, but not a morally wrongful one. In contrast, I suggest that discrimination is always morally wrong or, to adopt Lippert-Rasmussen's phrase: An act is pro tanto an act of discrimination, if it is, because it wrongs a person belonging to a certain group by means of an inappropriate balancing of relevant ethical concerns. This is in line with a widespread use of the notion in everyday language (Thomsen 2013, note 15). It is also supported by anti-discrimination provisions in legal frameworks. Typically, such provisions prohibit discrimination *simpliciter* which, in turn, suggests that discrimination is always wrong.

To think of "discrimination" as necessarily rather than only contin gently bad does, of course, not settle the question, why discrimination is morally wrong. The unreflective use of the notion in everyday language often obscures this fact and the philosophical discussion of recent years has, therefore, raised the question about the wrongfulness of discrimination with good reasons. While most scholars have tried to identify one single wrong-making feature of discrimination, I suggest that there is no such single wrong-making feature. Rather, there are a number of different features that can play a role, among them, of course, harm (as argued by Lippert-Rasmussen), but also disrespect (as argued by Hellman) or infringements of deliberative freedoms (as argued by Moreau). The conditions (iii) and (iv) of the modified definition capture two characteristic features of discrimination: first, discrimination is always tied to group membership (though the group need not be "social salient") and second, discrimination is always linked to a balancing process. Note that this implies that a person who discriminates against someone else must have a prima facie moral reason for his acting (some reason to put in the balance). If, for example, someone brutally batters someone else because of a certain feature (i.e., group membership) we would hardly call this an act of discrimination, but rather plain battery. If, however, an employer refuses to give a job to an applicant because of a certain feature that is unrelated to the job we certainly would call this discrimination. In such a case, the employer can prima facie claim to have a right to give the job to whomever he or she wants. The applicant, in contrast, can claim to be disadvantaged. Only after a weighting of the competing claims we come to the conclusion that the right to assign the job without constraints is morally less weighty than the right of the applicant to be considered solely according to his or her job-related qualifications. All things considered, the right to assign the job without constraints turns out to be insufficient. Again, the latter is characteristic for discrimination: it is always an *all things considered*-issue or weighting of various ethical concerns.

## 5 An exemplary case

A good example to illustrate the complex interplay of different ethical concerns that are at stake when it comes to deciding whether an act should count as a case of discrimination or not is the recruitment practice of religious institutions.[2] For quite a long time, religious institutions could impose specific requirements on job applicants, especially in view of religious commitment—a trait which is typically explicitly mentioned in anti-discrimination provisions. Recently, this established practice has been restricted by a ruling of the European Court of Justice (case C-414/16 as of 17 April 2018). The court decided that

> "the genuine, legitimate and justified occupational requirement it [i.e. Art. 4(2) of Directive 2000/78] refers to is a requirement that is necessary and objectively dictated, having regard to the ethos of the church or organisation concerned, by the nature of the occupational activity concerned or the circumstances in which it is carried out, and cannot cover considerations which have no connection with that ethos or with the right of autonomy of the church or organisation. That requirement must comply with the principle of proportionality." (Rn. 69)

In other words, religious institutions—in this particular case, the Protestant Church in Germany and its social welfare organization Diakonie—are no longer allowed to reject job applicants with deviant religious beliefs as such, but only if the nature of the occupational activity concerned or the circumstances in which it is carried out have a strong connection to the ethos of the institution. This connection must, in principle, be subject to an independent re-examination and must not rest on the evaluation of the religious institution alone. In this particular case, the national labor court to which the case was redirected saw no such strong connection and considered, therefore, that the rejection of the applicant was unlawful (German Federal Labour Court, case 8 AZR 501/14 as of 25 October 2018). Yet, in another case, the German Federal Constitutional Court supported the right of religious communities—in this case, the Catholic Church

[2] In this paper, I am primarily concerned with the *ethical* concept of discrimination. *Legal cases* like those I am considering in the following can, therefore, not count as decisive evidence for my argument. It might be that existing legal regulations simply do not measure up to the ethical concept of discrimination. Moreover, court decisions could be based on various considerations or even employ a particular legal concept of discrimination different from the ethical one. However, the ethical discussion on discrimination of recent years which I am taking up is largely guided by an established use of the concept of discrimination which, in turn, is reflected (among other things) in legal regulations and court decisions. Considering them may, therefore, at least provide some evidence.

and its social welfare organization Caritas—to set its own standards for employees. In this case, a chief physician was released from his position in a catholic hospital after his remarriage (case 2 BvR 661/12 as of 22 October 2014). Apparently, the court did not count this as a case of discrimination. The German Federal Labour Court forwarded this decision to the European Court of Justice, which, again, decided against the religious community (case C-68/17 as of 11 September 2018). The rulings by the European Court of Justice reflect a certain dynamic in view of the social consensus regarding the authority of religious communities on the one side and the autonomy of the individual on the other. According to the new weighting authorized by the court, what did not count as discrimination so far is considered a discriminatory practice as of now. The ruling by the German Federal Constitutional Court shows, in contrast, that this new weighting is not (yet) generally accepted. For the time being, the notion of discrimination is somewhat indeterminate in this context.

Some may wish to object that the case of religious communities is a very special one that is hardly suitable to demonstrate the dynamic character of the concept of discrimination. There are, however, other examples which point along similar lines. Take, for instance, the decision by the European Court of Justice that gender-related differences in insurance contracts are incompatible with European law (case C-236/09 as of 1 March 2011).

All this confirms, again, that social saliency is not suitable to define discrimination. Rather, it is always a mixture of different normative considerations that merge into the concept of discrimination. What is more, there is no globally applicable standard that determines the right weighting. While in one social environment, a trait may be considered especially important (e.g., because of specific historical injustice or present disadvantages), it can be less critical in another and may, therefore, more easily be dismissed. While individual freedom is considered paramount in one context, protection from harm for members of a certain group is more important in another.

If the concept of discrimination is used in an unqualified way, it is often suggested that there is an established weighting of the various ethical concerns involved. Sometimes such an established weighting does really exist, and the concept of discrimination can rightly be used to flag it. Yet, sometimes such established weighting does not exist. Then substantial reasons need to be provided to reach it. Simply stating that a certain act is a case of discrimination is insufficient under these circumstances and merely question begging.

## 6 The rise of AI/ADM as a matter of ethical concern

Since decades ago, a large variety of computer-based tools have been employed and continually refined to support decision-making in many areas of human life. In view of this development, Virginia Eubanks has observed:

> "Since the dawn of the digital age, decision-making in finance, employment, politics, health, and human services has undergone revolutionary change. Forty years ago, nearly all of the major decisions that shape our lives—whether or not we are offered employment, a mortgage, insurance, credit, or a government service—were made by human beings. They often used actuarial processes that made them think more like computers than people, but human discretion still ruled the day. Today, we have ceded much of that decision-making power to sophisticated machines. Automated eligibility systems, ranking algorithms, and predictive risk models control which neighborhoods get policed, which families attain needed resources, who is short-listed for employment, and who is investigated for fraud." (Eubanks 2017, 3)

Shortly before, Frank Pasquale has subsumed different technologies under the critical notion of "secret algorithms" and proclaimed the "black box society" (Pasquale 2015). In a more popular approach, Cathy O'Neil has coined the term "weapons of math destruction" to emphasize the dramatic consequences of AI/ADM on society (O'Neil 2016). Meanwhile, a number of scholarly papers address "the ethics of algorithms" (Mittelstadt et al. 2016).

For some time, everybody has been talking about artificial intelligence (AI). However, the notion is pretty vague. Famously, Elaine Rich, Kevin Knight and Shivashankar Nair provided the following—admittedly tentative—definition:

> "*Artificial Intelligence* (AI) is the study of how to make computers do things which, at the moment, people do better." (Knight et al. 2010, 3)

Providing a precise definition proves to be difficult and is still a matter of scientific dispute (cf. Bringsjord and Govindarajulu 2020). Within a public policy context, the EU High-Level Expert Group on Artificial Intelligence suggested the following working definition:

> "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and

> deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)." (High-Level Expert Group in Artificial Intelligence 2019a, 6)

Apparently, the kind of systems for automated decision-making currently discussed largely overlaps with this notion of AI. At times, an even broader notion of AI serves as a convenient label that subsumes all sorts of computer-based tools. The ongoing discussion on the "ethics of algorithms" is partly based on such a broader notion of AI. For the moment, this somewhat unspecific talk of AI will be taken as a basis. Further below, some qualifications will be made.

## 7 Discrimination and AI/ADM

A recurrent theme within the criticism of AI/ADM is that its increasing use aggravates issues of discrimination. For example, Eubanks remarks:

> "Automated decision-making shatters the social safety net, criminalizes the poor, intensifies discrimination, and compromises our deepest national values. It reframes shared social decisions about who we are and who we want to be as systems engineering problems." (Eubanks 2017, 12)

In a latter passage, she observes:

> "High-tech tools have a built-in and patina of objectivity that often lead us to believe that their decisions are less discriminatory than those made by humans. But bias is introduced through programming choices, data selection, and performance metrics. The digital poorhouse, in short, does not treat like cases alike." (Eubanks 2017, 194–195)

From the context, it becomes clear that Eubanks thinks that AI/ADM is not only no better than human decision-making, but much worse.

In addition, Pasquale has, among other things, discriminatory practices in mind when he expresses his concerns in view of the widespread use of AI/ADM. According to him, "[w]ithout a society-wide commitment to fair data practices, digital discrimination will only intensify." (Pasquale 2015, 21).

Such charges are not limited to popular books or semi-academic literature. In a scholarly review paper, Mittelstadt et al. remark:

> "Much of the reviewed literature also addresses how discrimination results from biased evidence and decision-making." (Mittelstadt et al. 2016, 8).

In fact, in legal theory, philosophy, the social sciences and computer sciences alike, the issue of "discrimination in the age of algorithms" (Kleinberg et al. 2018) has become an intensively debated topic (cf. e.g. Zliobaite 2017; Danks and London 2017; Chander 2017; Kim 2017; Gillis and Spiess 2018; Lee 2018; Williams et al. 2018; Gangadharan and Niklas 2019; Obermeyer et al. 2019).

In the meantime, the issue has also reached public policy debates (Gómez 2018, chap. 8; Global Future Council on Human Rights 2018). Zuiderveen Borgesius has written a report for the Anti-Discrimination Department of Council of Europe. In this report, the author claims that non-discrimination law has "several weaknesses in the context of AI decision-making" (Zuiderveen Borgesius 2018, 19). Moreover, Zuiderveen Borgesius maintains that "new types of AI-driven differentiation seem unfair and problematic – some might say discriminatory" (Zuiderveen Borgesius 2018, 20). However, why exactly is the use of AI/ADM problematic in the first place? In addition, why does it intensify discrimination? Some would argue exactly the other way around, as Eubanks willingly reports:

> "In Indiana, Los Angeles, and Allegheny County, technologists and administrators explained to me that high-tech tools in public services increase transparency and decrease discrimination." (Eubanks 2017, 168)

More generally speaking, how is it possible that in so many areas AI/ADM is being introduced if it is morally problematic? Do people advocating its use simply ignore the problems or are they suffering from moral turpitude? If it would be undisputable that AI/ADM aggravates discrimination, then the latter would be the only reasonable conclusion.

In a short, but illuminating analysis on "Data Mining and the Discourse on Discrimination" Solon Barocas has suggested that "[i]nconsistencies in charges of discrimination have been the cause of recent teeth gnashing" and has continued to remark that "the current debate suffers from many of the same conceptual challenges that have characterized discrimination from its very inception as a formal notion in the law" [and in ethics, as might be added] (Barocas 2014, 3–4). The analysis provided above may help to elucidate the issue. If the argument there was sound, the first question to ask is whether there is an established weighting of relevant

ethical principles for a specific area or not. There are three possible answers to this question which give rise to three different scenarios: (1) an established weighting of relevant ethical principles does not exist, i.e., the concrete reading of discrimination is unsettled, (2) such a weighting does exist in principle, but its exact scope or mode of application is contentious in detail, and (3) a weighting does exist, i.e., the notion of discrimination is definitive. I will examine the claim that the use of AI/ADM aggravates issues of discrimination for these three scenarios in turn.

(1) If there is no established weighting of relevant ethical principles for a specific area, substantial reasons need to be provided to reach it. Only then, it is possible to classify a certain treatment as discriminatory—regardless of whether AI/ADM is involved or not. In this case it is, therefore, inappropriate to maintain that the use of AI/ADM aggravates issues of discrimination. It may, rather, trigger a debate on how to weigh relevant ethical principles that is overdue.

(2) In some cases, anti-discrimination rules exist, but their scope or appropriate mode of application is contentious. This is, apparently, the problem in some of the controversies about AI/ADM. An interesting case in point has recently been reported from Austria's employment agency (Arbeitsmarktservice—AMS).

According to a report from Algorithm Watch (Kayser-Bril 2019), AMS is about to roll out an algorithm that assigns job seekers to three different categories based on their estimated chances on the labor market and the predicted benefit of support measures. Persons who probably need no help in finding a new job are assigned to category A, persons who might benefit from retraining are assigned to category B, and persons who are deemed unemployable are assigned to group C. The sorting is based on a number of personal traits, including sex, age, nationality, education, and health status. The reason for introducing the algorithm is that the employment agency wants to spend its limited resources most efficiently and seeks to avoid spending money for support measures which are futile. However, a documentation paper by the company that designed the algorithm (Holl et al. 2018) shows that members of some groups, e.g., women and disabled persons, are given a negative weight by the algorithm in at least some models. Critics like Algorithm Watch, therefore, claimed that the algorithm is discriminatory. AMS executive director Johannes Kopf rejected such objections as mistaken and unfounded (Kopf 2019). He reaffirmed that the algorithm is meant as a means for improving the decisions taken by the agency's consultants.

Apparently, there is a dispute about the appropriate way of applying existing anti-discrimination rules between AMS and its critics. While the critics see a clear violation of existing national and EU anti-discrimination rules, the agency seems to argue that spending public money inefficiently is wrongful and unfair against those who would benefit from supporting measures. In part, this is a dispute about whether (or to what degree) past injustices and present disadvantages should be included in the weighting of ethical principles. As has been argued above, such concerns typically play an important role in the process of determining what has to count as discrimination. Yet, the specific weight which these concerns should be given is sometimes controversial. To sum up, the use of AI for decision-making does not aggravate issues of discrimination in this second scenario. Again, it only highlights the need for an ethical debate—this time about the correct interpretation of the concept in a given context. In this debate, neither side can merely claim that issues of discrimination are at stake or reject this claim, respectively. As long as there is no agreement on moral grounds, the question of discrimination is simply open.

(3) Existing anti-discrimination rules already cover a broad field of activities, in particular, many of those in which AI/ADM is already being used or will be used in the near future. On the face of it, it should be clear in these cases whether a certain algorithm is discriminatory or not. Again, the charge of aggravation would be mistaken. However, things are more complicated as I will show in the next section.

## 8 Discrimination intensified

As has been mentioned above, Pasquale deploys the notion of a "black box" in his critique of the use of AI for decision-making. In doing so, he points towards something important. What he highlights is a lack of transparency that is typical for many uses of AI/ADM (Pasquale 2015, 3–14). In fact, AI/ADM is often characterized by "epistemic opacity", that is to say, the algorithms include epistemically relevant elements which a cognitive agent does not or even cannot know (Humphreys 2009). The complexity of AI/ADM makes it difficult to evaluate its inner workings. What is more, the ability of machine-learning approaches to derive complex patterns of high-dimensional interactions lies at the heart of their power and renders them non-transparent. The outcome they generate often does not allow for unpacking and accessing its detailed formation (Heinrichs and Eickhoff 2020).

Epistemic opacity is the real problem when it comes to automated decision-making and the charge of discrimination. If we need to come to a shared understanding regarding an appropriate weighting of competing ethical concerns to determine whether something should count as discrimination or not, we need to know what parameters were included in the first place, how the data were collected, and how it influenced the outcome. Epistemic opacity undermines exactly this process and *therefore* intensifies issues of discrimination. Or, to put it differently, AI/ADM often is at odds with our common practice of ethical deliberation. Since

the notion of discrimination rests on exactly this practice of ethical deliberation, it is threatened by the usage of these forms of automated decision-making.

Above, I have distinguished three cases based on whether an established weighting of ethical principles exists that defines what has to count as discriminatory practice in a particular area. If such a weighting does not exist or its precise scope or appropriate mode of application is contentious, the claim that the use of AI/ADM turned out to be erroneous. I will now consider the case that an established weighting of principles does exist. In fact, the prominent role which the concept of discrimination plays in modern societies has given rise to a wide range of anti-discrimination rules and a sophisticated casuistry of their proper application. Quite often, cases of discrimination are not difficult to detect and regulatory mechanisms are well suited to sanction them. However, the use of AI/ADM can undermine this well-functioning normative practice. In short, the claim that the use of AI/ADM aggravates issues of discrimination is in some cases accurate. In the following, I will examine three different scenarios which all have a central theme: a lack of transparency.

First, I want to examine cases in which it is unclear exactly which parameters are used in an algorithm. An instructive example is the "Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)" tool originally developed by Northpointe (now distributed by Equivant as part of the Northpointe Suite) and used by courts in several US states. Among other things, the system is being used for rating a defendant's risk of future crime. Severe criticism has been raised by ProPublica that COMPAS is biased against blacks (Angwin et al. 2016). Equivant disagreed and claimed that ProPublica made several statistical and technical errors (Dieterich et al. 2016). Others have supported their view (Flores et al. 2016). So far, it is impossible to assess the accusation as well as the apology, since Equivant does not fully disclose the algorithm used to calculate a defendant's risk score. For the time being, it remains unclear whether COMPAS is actually discriminatory or not (conceding that an agreement about the appropriate weighting of ethical principles exists). Considering the importance of the area in which COMPAS is used, this is hardly acceptable.[3]

A second type of case is equally important in view of discriminatory practices and also linked to epistemic opacity: the use of proxies for protected personal traits. It is a common method of bypassing existing anti-discrimination rules to use a proxy trait instead of an explicitly protected trait.[4] Say, it is agreed that gender is a trait that must not be considered for selecting persons in a particular context. It is, then, possible to base a selection procedure on another accessible trait X which is highly correlated with gender, but not explicitly protected rather than on gender itself with the same discriminatory effect. The crucial point is that the use of AI/ADM may facilitate this type of discrimination and, at the same time, obscure it, as Mittelstadt points out:

> "Proxies for protected attributes are not easy to predict or detect […], particularly when algorithms access linked datasets […]. Profiles constructed from neutral characteristics such as postal code may inadvertently overlap with other profiles related to ethnicity, gender, sexual preference and so on […]. Beyond legally protected groups. It remains unclear from the outset the types of behavioural identity tokens and decision-making models that can be produced, and which of these are potentially ethically troubling." (Mittelstadt 2017, 479).

Large datasets make it easy to identify suitable proxies while the epistemic opacity of AI/ADM algorithms makes it difficult to access their exact method of operation. As a consequence, the use of AI/ADM can really aggravate issues of discrimination for it can undermine the enforcement of established ethical agreements. Barocas puts forward that in such cases "[…] data mining is not itself discriminatory. Rather, data mining here serves as a tool for those who purposefully seek out new ways to discriminate." (Barocas 2014, 1) He then admits that "data mining can more effectively realize discriminatory intent"—which I take to be a form of aggravation. However, Kleinberg et al. (2018) maintain that algorithms have also the potential for increasing transparency. In particular, they emphasize that human decisions are often opaque, too, and that algorithms may help to detect forms of discrimination that otherwise would remain hidden or at least unproven.

A third type of cases concerns the gathering of (training) data which is essential for many AI/ADM technologies. In a helpful review, David Danks and Alex John London distinguish five types of algorithmic bias (Danks and London 2017, 4692–4694). In view of the first type examined, i.e., "training data bias", they have observed:

> "In particular, a 'neutral' learning algorithm (in whatever sense of that term one wants) can yield a model that strongly deviates from the actual population statistics, or from a morally justifiable type of model, simply

[3] It should be noted that the intense debate about the COMPAS algorithm is primarily concerned with a different question: the point at issue is whether the algorithm's aggregate effect on particular groups is morally acceptable or not.

[4] Arguably, this makes it a case of indirect rather than direct discrimination. It is controversial whether these two types of discrimination are morally different or not, cf. Khaitan (2018).

> because the input or training data is biased in some way. Moreover, this type of algorithmic bias (again, whether statistical, moral, legal, or other) can be quite subtle or hidden, as developers often do not publicly disclose the precise data used for training the autonomous system. If we only see the final learned model or its behavior, then we might not even be aware, while using the algorithm for its intended purpose, that biased data were used." (Danks and London 2017, 4692).

There are a number of well-documented real cases and also fictional ones which demonstrate the subtlety the authors here refer to. An especially interesting case has recently been reported by Obermeyer et al. (2019) They investigated a widely used algorithm for managing the health of populations. According to their analysis, the algorithm "exhibits significant racial bias" (Obermeyer et al. 2019, 447). The reason for this bias is not at all obvious and was only detectable because the researchers had "a unique window into the mechanisms by which bias arises." (Obermeyer et al. 2019, 449). They found that "the algorithm's prediction on health needs is, in fact, a prediction on health costs." (Obermeyer et al. 2019, 449). However, it turned out that at "a given level of health […] Blacks generate lower costs than Whites […]." (Obermeyer et al. 2019, 450) The authors conclude:

> "These results suggest that the driving force behind the bias we detect is that Black patients generate lesser medical expenses, conditional on health, even when we account for specific comorbidities. As a result, accurate prediction of costs necessarily means being racially biased on health." (Obermeyer et al. 2019, 450)[5]

This case shows, once again, that the use of AI/ADM can aggravate issues of discrimination—even unintentionally. The inner workings of AI/ADM itself can have a detrimental effect and hinder transparency which is essential for avoiding discrimination.

## 9 A call for transparency and explainability

Given the "ethical grammar" of the notion of discrimination, its proper use—flagging an agreed-upon malpractice—is already difficult and sometimes not adequately observed. With the introduction of AI/ADM, it can become more difficult, and sometimes even impossible, to apply the notion correctly—at least if we do not impose strict transparency requirements on the implementation and use of AI/ADM. For an ethical assessment, it is essential to know details about algorithms. This includes knowledge about technical details such as target variables and class labels, training data and feature selection (Barocas and Selbst 2016, 678 ff.).

The EU High-Level Expert Group on Artificial Intelligence rightly emphasizes the principle of explicability in their recent *Ethics Guidelines for Trustworthy AI*:

> "Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate." (High-Level Expert Group on Artificial Intelligence 2019b, 13)

In this passage, the notion of "black box" is used in a more limited sense than in Pasquale's book. The High-Level Expert Group on AI employs it to mark important differences within the broad field of AI/ADM technologies. These differences are important since they give rise to different transparency requirements. It is, therefore, now the time to make some distinctions in view of AI/ADM. In some cases, systems for automated decision-making use rather simple models. In these cases, it is essential to know details about the data and the target variables. In principle, transparency is easy to realize in such cases—unless companies or organizations are blocking the disclosure of their algorithms—although the amount of data can still be a challenge in such cases. In the case of more complex models (in particular in case of complex deep learning algorithms), creating transparency is much more difficult. In these cases, it is not the reluctance of a company or organization or the complexity of correlations that prevents openness, but the nature and design of the algorithms themselves. While in view of simple types of automated decision-making, it can already be difficult to access which features they appeal to explicitly or implicitly, complex algorithms (e.g., deep learning

[5] This case could also be considered under the second scenario above, i.e., as a problem of inappropriate proxies. In considering it here, I want to emphasize the discrimination-relevant effect of training data on the outcome an algorithm generates.

algorithms) proceed in ways that can be inaccessible for human reasoning.

Under the notion of "explainable AI", efforts are being made to address the aforementioned problem (cf. Wierzynski 2018). Measures such as traceability, auditability and transparent communication on system capabilities are perhaps the best that can be done to make complex algorithms accessible for critical evaluation. It turns out that AI is a heterogenous field: while some forms of AI are only "contingently opaque" (e.g., due to trade secrets) others are "essentially opaque" (i.e., due to the inherent complexity of the models used). This must be taken into account when calling for transparency (Selbst and Barocas 2018). At any rate, appropriate ways need to be developed to make algorithms intelligible.

In view of the problem of discrimination, transparency regarding the use of personal traits is crucial. What is needed, first, is an open debate about which traits must not be used for differentiating between people in specific contexts. The agreements reached in this debate are the basis for critically evaluating the use of AI/ADM and developing appropriate tools (see for example Zliobaite 2017). It might be that we conclude that in specific contexts the use of particular traits is incompatible with our vision of a fair society. If we do, we need to know whether an algorithm makes use of them or not and how they influence its output. It may, therefore, only superficially be a paradox to ask for more data to fight discrimination in the age of algorithms (Williams et al. 2018). However, acquiring and using data in an appropriate way becomes certainly more challenging. In this sense, issues of discrimination really are intensified in the age of AI/ADM.

## 10 Conclusion

In this paper I have examined the claim that the use of artificial intelligence and automated decision-making (AI/ADM) aggravates issues of discrimination. I have rejected this general claim on the basis of a philosophical understanding of discrimination. In contrast to some other accounts, I have defended a moralized understanding of discrimination. What is more, I have argued in favor of the formula: "Action A is a case of discrimination, *because of* ω [where ω refers to an established weighting of relevant ethical concerns in a given context]". This account helps to understand why the general claim in view of the aggravation of discrimination is unwarranted. However, epistemic opacity may undermine a sound ethical examination of complex algorithms and this, in turn, can intensify issues of discrimination. Against this background, research initiatives for explainable AI are especially important from an ethical point of view.

**Acknowledgements** I would like to sincerely thank my colleagues at INM-8 as well as at IWE for comments on an earlier version of this paper. My special thanks go to Annette Dufner and Sebastian Knell, who pointed out a number of weaknesses in my argumentation. I hope that I have been able to clear them up, at least in part. Finally, I would like to thank Charles Rathkopf for his assistance with the final editing as well as Ulrich Steckmann and Sandra Fömpe super fast proofreading.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

Alexander L (1992) What makes wrongful discrimination wrong? Biases, preferences, stereotypes and proxies. Univ Pa Law Rev 141:149–219. https://doi.org/10.2307/3312397

Altman A (2020) Discrimination. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (summer 2020 edition). https://www.plato.stanford.edu/archives/win2016/entries/discrimination/. Accessed 11 June 2019

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 14 June 2019

Arneson RJ (2006) What is wrongful discrimination? San Diego Law Rev 43:775–808

Barocas S (2014) Data mining and the discourse on discrimination. In: Proceedings of data ethics workshop. https://www.dataethics.github.io/proceedings/DataMiningandtheDiscourseOnDiscrimination.pdf. Accessed 11 June 2019

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif Law Rev 104:671–732. https://doi.org/10.2139/ssrn.2477899

Bringsjord S, Govindarajulu NS (2020) Artificial intelligence. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (Summer 2020 Edition). https://www.plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/. Accessed 25 March 2021

Chander A (2017) The racist algorithm? Mich Law Rev 115:1023–1045

Cook R (2015) Discrimination revised: reviewing the relationship between social groups, disparate treatment, and disparate impact. Moral Philos Polit 2:219–244. https://doi.org/10.1515/mopp-2014-0026

Danks D, London AJ (2017) Algorithm bias in autonomous systems. In: Proceedings of the twenty-sixth international joint conference on artificial intelligence (IJCAI-17), pp 4691–4697

Dieterich W, Mendoza C, Brennan T (2016) COMPAS risk scales: demonstrating accuracy equity and predictive parity performance of the COMPAS risk scales in Broward County. http://www.go.

volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf. Accessed 15 July 2019
Eubanks V (2017) Automating inequality. How high-tech tools profile, police, and punish the poor. St. Martin's Press, New York
European Union Agency for Fundamental Rights (2011) Handbook on European non-discrimination law. Publications Office of the European Union, Luxembourg
Ezorsky G (1992) Discrimination. In: Becker LC (ed) Encyclopedia of ethics, vol 1. St. James Press, Chicago, pp 264–267
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 80:38–46
Gangadharan SP, Niklas J (2019) Decentering technology in discourse on discrimination. Inf Commun Soc 22:882–899
Gillis TB, Spiess JL (2018) Big Data and discrimination. Harvard John M. Olin fellow's discussion paper series 84
Global Future Council on Human Rights (2018) How to prevent discriminatory outcomes in machine learning. White Paper, Geneva
Gómez E (2018) (ed) Assessing the impact of machine intelligence on human behavior: an interdisciplinary endeavor. In: Proceedings of the 1st HUMAINT workshop, Barcelona, Spain, March 5–6, 2018, Luxemburg
Halldenius L (2005) Dissecting "discrimination." Cambr Q Health Care Ethics 14:455–463. https://doi.org/10.1017/s0963180105050619
Heinrichs B (2007) What is discrimination and when is it morally wrong? Jahrbuch für Wissenschaft und Ethik 12:97–114. https://doi.org/10.1515/9783110192476.1.97
Heinrichs B, Eickhoff SB (2020) Your evidence? Machine learning algorithms for medical diagnosis and prediction. Hum Brain Mapp 41:1435–1444
Hellman D (2008) When is discrimination wrong? Harvard University Press, Cambridge
High-Level Expert Group in Artificial Intelligence (2019a) A definition of AI: main capabilities and disciplines. https://www.ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341. Accessed 17 June 2019
High-Level Expert Group in Artificial Intelligence (2019b) Ethics guidelines for trustworthy AI. https://www.ec.europa.eu/newsroom/dae/document.cfm?doc_id=58477. Accessed 17 June 2019
Holl J, Kernbeißer G, Wagner-Pinter M (2018) Das AMS-Arbeitsmarktchancen-Modell. Dokumentation zur Methode. Synthesis-Forschung, Wien. https://www.ams-forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf. Accessed 25 March 2021
Humphreys P (2009) The philosophical novelty of computer simulation methods. Synthese 169:615–626. https://doi.org/10.1007/s11229-008-9435-2
Kayser-Bril N (2019) Austria's employment agency rolls out discriminatory algorithm, sees no problem. https://www.algorithmwatch.org/en/story/austrias-employment-agency-ams-rolls-out-discriminatory-algorithm/. Accessed 11 June 2019
Khaitan T (2018) Indirect discrimination. In: Lippert-Rasmussen K (ed) The Routledge handbook of the ethics of discrimination. Routledge, London, pp 30–41
Kim PT (2017) Data-driven discrimination at work. William & Mary Law Rev 58:857–936
Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR (2018) Discrimination in the age of algorithms. J Legal Anal 10:113–174
Kopf J (2019) Ein kritischer Blick auf die AMS-Kritiker. Der Standard. https://www.derstandard.de/story/2000109032448/ein-kritischer-blick-auf-die-ams-kritiker. Accessed 11 June 2019
Lee NT (2018) Detecting racial bias in algorithms and machine learning. J Inf Commun Ethics Soc 16:252–260
Lippert-Rasmussen K (2006) The badness of discrimination. Ethical Theory Moral Pract 9:167–185. https://doi.org/10.1007/s10677-006-9014-x
Lippert-Rasmussen K (2013) Born free and equal? A philosophical inquiry into the nature of discrimination. Oxford University Press, New York
Lippert-Rasmussen K (ed) (2018) The Routledge handbook of the ethics of discrimination. Routledge, London
Mittelstadt B (2017) From individual to group privacy in big data analytics. Philos Technol 30:475–494
Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc 3:1–21. https://doi.org/10.1177/2053951716679679
Moreau S (2010) What is discrimination? Philos Public Aff 38:143–179. https://doi.org/10.1111/j.1088-4963.2010.01181.x
Nickel JW (1998) Discrimination. In: Craig E (ed) Routledge encyclopedia of philosophy, vol 3. Routledge, London, pp 103–106
O'Neil C (2016) Weapons of math destruction. how big data increases inequality and threatens democracy. Crown, New York
Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366:447–453
Pasquale F (2015) The Black Box Society. The secret algorithms that control money and information. Harvard University Press, Cambridge
Knight K, Rich E, Nair SB (2010) Artificial intelligence, 3rd edn. Tata-McGraw-Hill, New Delhi
Segall S (2012) What's so bad about discrimination? Utilitas 24:82–100. https://doi.org/10.1017/s0953820811000379
Selbst AD, Barocas S (2018) The intuitive appeal of explainable machines. Fordham Law Rev 87:1085–1139. https://doi.org/10.2139/ssrn.3126971
Thomsen FK (2013) But some groups are more equal than others—a critical review of the group-criterion in the concept of discrimination. Soc Theory Pract 39:120–146. https://doi.org/10.5840/soctheorpract20133915
Thomsen FK (2018) Direct discrimination. In: Lippert-Rasmussen K (ed) The Routledge handbook of the ethics of discrimination. Routledge, London, pp 19–29
Wasserman D (1998) Discrimination, concept of. In: Chadwick R (ed) Encyclopedia of applied ethics, vol 1. Academic Press, San Diego, pp 805–814
Wierzynski C (2018) The challenges and opportunities of explainable AI. https://www.ai.intel.com/the-challenges-and-opportunities-of-explainable-ai/. Accessed 5 Apr 2019
Williams BA, Brooks CF, Shmargad Y (2018) How algorithms discriminate based on data they lack: challenges, solutions, and policy implications. J Inf Policy 8:78–115
Zliobaite I (2017) Measuring discrimination in algorithmic decision making. Data Min Knowl Disc 31:1060–1089. https://doi.org/10.1007/s10618-017-0506-1
Zuiderveen Borgesius F (2018) Discrimination, artificial intelligence, and algorithmic decision-making. Strasbourg. https://www.rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73. Accessed 11 June 2019

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.