Racism and Artificial Intelligence

Katharina Ottilie Buch

Artificial Intelligence (AI) has become quite a buzzword over the last years. This comes as no surprise since AI increasingly penetrates companies, institutions, and everyday life – and therefore reaches new levels of influence. Research on the ethics of AI and forms of human bias in AI broadly is carried out in various fields: Science and Technology Studies (STS), philosophy of technology, and computer sciences are the fields that have primarily tackled these issues. Additionally, scholarly research has broadened immensely in the last years to include research on law, economics, labour, public policy, (urban) geography, sociology in general, and many other fields (Dwivedi et al. 2019; Franklin 2014). The research that has been conducted on racism and AI inhabits a niche inside this scope of scholarly research.

In this contribution, a general understanding of what AI is and why AI is such a debated topic will be established. This is fundamental to understanding how and why racism intersects with AI, especially due to so-called racial bias. How and why racial bias creeps into AI can then be explained. Examples of racial bias in AI illustrate this phenomenon across sectors including justice, policing, and credit scores, to healthcare, job markets, and consumer goods. In a next step, solutions that have been proposed in literature are presented. These include: technical solutions that involve re-building algorithms, data diversification, and transformation; solutions including regulations, transparency of AI, and education of AI developers/researchers, and practitioners working with AI, to diversification of AI research/development teams. In a final step, open questions and resulting research needs will be presented.

Understanding AI

Definition of AI

First, a good starting point to grasp AI is the broad definition provided by *The Cambridge Handbook of Artificial Intelligence*, which states that »[…] artificial intelligence (AI) is a cross-disciplinary approach to understanding, modelling, and replicating intelligence and cognitive processes by invoking various computational, mathemat-

ical, logical, mechanical, and even biological principles and devices« (Frankish & Ramsey 2014).

AI can be seen as an umbrella term that covers concepts such as knowledge-based expert systems, machine learning/deep learning, artificial neural networks, natural language processing, cognitive computing, and developmental robotics (Franklin 2014). Hence, when we talk about AI it can mean any of the mentioned systems, a combination of them, or a modification to them.[1] Additionally, the term ›algorithm‹ is commonly used in research papers, especially on human bias. However, ›algorithm‹ is a vague descriptor, as algorithms have been around for a much longer time than the AI systems these algorithms are usually part of. Hence, for coherence, the contribution will primarily use the terms AI and AI system.

Another distinction is to be made on the level of ›intelligence‹ of AI. The AI we see today is *narrow* AI. This means it can solve certain tasks, often more efficiently than humans. Its counterpart is Artificial *General* Intelligence (AGI). The concept of AGI is a complex one but at this point mainly speculative. There are many discussions to be had about its likelihood, complications, and ethics. For this paper, it suffices to say that the theoretical idea encompasses an AI that can solve (almost all) tasks better than humans and is able to abstract knowledge. Currently, while there is a lot of research on AGI, it is still only fiction. So, when talking about AI and its impacts on our lives today researchers usually mean *narrow* AI (Adams et al. 2012).

One more point to address is the level of autonomy of such AI. Levels of autonomy immensely impact how researchers, politicians, and citizens think and feel about its risks and threats. AI systems can be classified as AI with humans-in-the-loop, humans-on-the-loop, or humans-out-of-the-loop. Humans-in-the-loop refers to AI systems that are only deployed if a human agent requests or demands it. Humans-on-the-loop is AI that works autonomously – for example, by constantly sorting through data, identifying patterns, and automatically issuing recommendations. However, it is still a human agents who chooses whether and how to implement AI's recommendations. Humans-out-of-the-loop refers to an AI agent that acts and decides entirely on its own without human interference before (despite the development process), during, or after. The former two modes of AI are the ones commonly in practice today. Additionally, there seems to be a level of consensus, especially for AI in governmental settings, that humans should always ›have a say‹ in AI decision-making processes and, hence, humans-out-of-the-loop is seen as not desirable, at least for the time being (Campbell-Verduyn et al. 2017; Danaher 2016).

AI is still often seen as – at least potentially – value-neutral and more ›rational‹ than humans. However, it has already been shown many times that AI is not only

1 For a more comprehensive overview of what AI entails I suggest Stan Franklin's »History, Motivations, and core themes« (2009).

fallible but also able to inherit, display, and amplify multiply human biases, one of which is racial bias. Another debated topic is the accountability and transparency of AI. While this issue might not seem directly related to racism and AI, it is important to grasp, especially when one starts to think about solutions for ensuring that racial bias is not perpetuated.

The Black Box Problem

A commonly raised concern with AI is its opacity, its lack of transparency – its existence as a black box. Most famously, Frank Pasquale (2015) has shown how certain types of AI may lead to a lack of accountability and trustworthiness. This is mainly due to a lack of comprehensibility about how AI arrives at decisions for end-users, politicians, and civil society, sometimes even for experts.[2] In the case of complex neural networks, the problem of transparency is exacerbated because it is also difficult for the programmers themselves to understand how AI arrives at decisions. In the context of bias, this is of utter importance as one cannot fully pinpoint at which point AI ›failed‹ without being able to reconstruct how certain decisions/suggestions/actions of AI have been made (Campbell-Verduyn et al. 2017; Danaher 2016).

This raises concerns not only in connection with the expedient search for errors but also, more fundamentally, with regard to the potential for control, predictability, and accountability. While the black box issue is often framed around private-sector interests of commercial secrecy and in terms of competitiveness versus the interests of the end-user, the black box phenomenon can be particularly critical for the public sector since bureaucratic decision-making processes are explicitly designed to be traceable and should provide accountability through transparency (Erdélyi & Goldsmith 2018; Katyal 2019). In the chapter »Re-building Algorithms«, this issue will be explored more in-depth, as the chapter highlights proposed solutions for regulation and transparency by the scientific community and civil society.

Additionally, it is necessary to frame the issue of ›the black box AI‹ in a broader context. As hinted at before, AI is built, fed, and interpreted by humans. As we see more and more automation and technologically based predictions, we are faced with much more fundamental questions: What do we perceive as fair? How and how much must an agent – human or technological – explain itself to be considered accountable? Who is the audience – from experts to laypeople – that must be able to understand these decision-making processes? What epistemological school(s) of thought do we follow to understand the production of knowledge? There seems to be no way around these fundamental philosophical questions if we want to answer

2 It should be noted that the General Data Protection Regulation (GDPR) states that, in theory, people have a right to know why an algorithmic decision was made.

honestly and cohesively the question of how non-racist AI can be built (Beck 2020; Jones 2020). This, however, is a discussion beyond this contribution.

Overall, there is a very high likelihood that a society that discriminates – on the basis of race/ethnicity, gender, social class – will produce systems and technologies that carry on these prejudices (MacKenzie & Wajcman 1999). The higher standard that we expect from AI-driven predictions, and particularly from AI-driven decision-making, reveals an inherent bias towards human decisions. This is not necessarily a critique[3] but rather a reminder that current, more human-based modes of prediction, decision-making, and the like are all biased in one way or the other as well (Beck 2020; Kearns & Roth 2019). Hence, it comes as no surprise that there is a part of the scientific community that advocates for AI as a tool to combat human racial bias/racism. This idea will be tackled in the chapter »AI as Part of the Solution«.

Bias in AI

As hinted at before, AI can be biased. Bias refers to prejudice that is expressed through or built into the technology. Overall, it is important to understand that the incorporation of bias is usually unintentional and very human. Most researchers agree that knowledge is fundamentally perspectival. Hence programmers perceive the world through their specific position, through their cultural and social framework. This means AI systems are built by value-driven people whose norms, morals, and ethics can – subconsciously – influence their construction and design. Additionally, AI is trained with data sets that are not value-neutral. These data sets may reinforce stereotypes through ›unbalanced‹ distributions or rely on historical data that is inherently shaped by past discrimination (Kleinberg et al. 2018).

Differences Between Algorithmic Bias and Data Bias

Most problems with bias in AI arise inside the code or the data. Those two fields can be split into *algorithmic bias* – bias that arises from the algorithm, the technical AI design – and *data bias* – bias that arises from flawed and/or limited data sets. Algorithmic bias is usually based on either human biases creeping in or a lack of understanding of the context that algorithms require (Jones 2020; Zarsky 2016). An example is facial recognition software that has learned to recognise faces based on contrasts between different parts of the face; however, this assumption is based on

3 It may be noted that there are scholars critiquing such a human bias. It is, however, unfeasible to expect any kind of science to move beyond our bias for human intelligence as, at least so far, there has not been away around this.

how white faces are recognised – the fact that this might not be true for black faces was not thought through by the creators (Buolamwini & Gebru 2018).

With data bias, the problem lies in erroneous, limited data sets. Data sets serve as the learning basis for many AIs to identify patterns and make predictions. Due to ›bad‹ data sets, AIs can come to assumptions that are seen as racist, sexist, ableist etc. One example is recruitment software for software developer jobs and the like that discriminated against female candidates: The data set was composed of previously successful candidates and most employees had been male up to this point. Hence, the software started penalising phrasings such as ›women's football coach‹ as well as women's-only colleges. Human biases can be reproduced and reinforced through technical design and biased data sets. Though mostly without malicious intent, prejudice, stereotyping, and discrimination creep in. This is particularly problematic as AI often has a self-reinforcing effect, hence, it may not only reproduce existing racial inequities but also amplify them (Benjamin 2019; Jobin 2017; Kearns & Roth 2019).

Racial Bias in AI

Moreover, AI systems that employ inherently human traits, such as speech, usually strive towards being perceived as neutral as possible. It has been shown that White-sounding voices are the one's most commonly perceived as neutral in the global north. This comes as no surprise, as it has been shown repeatedly that non-White people are inherently framed in terms of their ethnicity/race. Only White people have the option to be perceived as basically devoid of race. Their voices, their bodies, and their culture are the ›norm‹ in the global north and, therefore, go without saying. An example: The AI voices most hear in their car, in their homes, are usually just perceived as such – as a voice. Every day many of us go without recognising the voice as explicitly White, but it is. Imagine a Black-sounding voice AI. There is no doubt that it would be seen as a statement, as it would be actively recognised as such. At this moment, one might realise just how ›neutralised‹ being White is. As most of the AI that mirrors humans tries to be as unassuming, digestible, and palatable as possible, AI usually mirrors Whiteness (Benjamin 2019; Cave & Dihal 2020).

Examples of racial biases and assumptions of ›White-as-default‹ in AI research, development, and deployment are manifold but often vague. When one seeks to analyse racial bias in AI in a sufficiently empirical manner, one runs into issues of publicly available data. As many AI systems are developed and deployed by private entities, the used models and data sets might not be available to the public and/or researchers. Hence, empirical analyses of those AI systems are often reliant on what companies readily disclose (Raghavan et al. 2019). The issue of transparency will also be addressed in the chapter »Proposed Solutions in Literature«. But even though data is limited, especially when it comes to private enterprises, racial bias in AI has

been reported in many areas, as the examples in the following chapters will highlight.

Racial Bias in … Justice

Racial bias exists in most justice systems, especially in the US, where most research has been conducted on the influence of AI in justice systems. The most commonly talked about AI systems in this area are prediction systems for recidivism, i.e., COMPAS in the US. Here, many studies have highlighted that predictions often have disproportionately negative effects on ethnic minorities – especially Black people (Berk et al. 2018; Skeem & Lowenkamp 2016). This comes as no surprise when one understands that predictions are primarily made based on historical data, which is skewed towards incarcerating Black people due to longstanding racial discrimination in the US (Chiao 2019). While these AI systems are not in use in Germany, one can still learn from case studies about them since they very might well be used in Germany at some point as well. Additionally, the disproportionate targeting of ethnic minorities in justice systems (and policing) is not limited to the US but is rather something we tend to see in most countries of the global North to different degrees and towards different minorities.[4]

In the US, AI prediction systems for recidivism are primarily used to prepare decisions for judges, either for bail or for early release. The final decision on each case remains to be made by a human judge (in research on AI this integration of humans is often labelled as ›human-in-the-loop‹). When evaluating the risk of recidivism, we are always faced with a dilemma: Not every person classified as ›high risk‹ might re-offend, not every person classified as ›low risk‹ might not. There are always trade-offs in accuracy – either by keeping people in prison that would have not re-offended (false-positives in the following) or by letting out people that re-offend (false-negatives) (Lupo 2019). The main issue is that AI prediction systems for recidivism tend to produce more false-positives for Black than for White people, as the former are often disproportionately grouped as ›high-risk‹. In short, there are more Black people that would not have re-offended that are kept imprisoned than there are White people that would have not reoffended (Angwin et al. 2016; Jiang 2020; Skeem & Lowenkamp 2016; Walsh et al. 2020). An interactive tool by MIT Technology Review (2019) highlights how those different levels of accuracy for different ethnic groups occur and what complicated calibrations developers and practitioners have to make.

Overall, the AI systems in place reinforce existing injustice, systemic issues, and inherent trade-offs that always have to be made but are now blatantly out in the open

4 An in-depth analysis of structural racism in justice and policing would go beyond this contribution.

(Arnold et al. 2020). While Germany does not use such systems, it is by no means a given that it will stay that way, as several AI prediction systems for recidivism have been tested in Europe (Lupo 2019). Furthermore, while current AI prediction systems for recidivism seem to be inherently flawed: there are arguments to be held that a consciously calibrated AI prediction system might produce less-biased results than a human judge alone.

… Policing

Another way in which AI comes more and more into play in the public sector – also in Germany – is the deployment of AI to analyse crime statistics and help develop prevention measures. Those systems are part of what is known as ›predictive policing‹. While not everything under the umbrella term ›predictive policing‹ relates to AI systems, many tools do since most use vast amounts of data and statistics to make predictions about where and what kind of crimes might occur (Egbert & Krasmann 2020).

Worldwide, legal scholars and urban geographers have focused on CompStat as the most commonly used data-driven, decision-making system for police forces internationally, including the US (Agrawal et al. 2019). In Germany, six federal states employ different prediction tools for policing (Seidensticker et al. 2018). The one prediction tool analysed most thoroughly is PRECOBS, which is in place in Munich and Middle Franconia and is being tested in Baden-Württemberg. It helps to identify high-risk areas for domestic burglary. Both CompStat and PRECOBS use the crime statistics of past years, which are based on police reports, not criminal behaviour – they are based on crimes either observed by police or reported by citizens. In the case of PRECOBS, however, only domestic burglary reports are taken into account (Egbert & Krasmann 2020; Sommerer 2017; Straube & Belina 2018). The prediction systems then make predictions based on the assumption that crimes that have been repeatedly reported in certain areas will more likely be reported again in these areas (Agrawal et al. 2019; Egbert & Krasmann 2020).

While this is a seemingly straightforward approach, there are two caveats. Firstly, this approach assumes that reported crimes are a reflection of reality. However, there are large numbers of unreported crimes that are fully neglected in these predictions. Secondly, it can become a vicious cycle: poorer communities with a higher rate of ethnic minorities (here, the intersection between social class and ethnicity comes strongly into play as well) are already more heavily policed due to higher crime rates in the past. By policing these areas more heavily, more crimes are witnessed and persecuted, leading to further stigmatisation of the area, and so on – we are stuck in a positive feedback loop (Safransky 2020). This predictive policing leads to disproportionate policing of poorer neighbourhoods and neighbourhoods with a large percentage of ethnic minorities. Hence, many scholars have raised con-

cerns that predictive policing systems ›naturalise‹ the racial bias that is produced through policing and governing urban space. The deployment of AI certainly has the potential to do nothing but manifest racist policing patterns and deepen disparities between neighbourhoods (Jefferson 2018; Safransky 2020; Shapiro 2017; Sommerer 2017).

However, these connections have been primarily shown in the US. Research on predictive policing in Germany does not thoroughly analyse connections between racism and predictive policing tools – especially not empirically (Hagendorff 2019). Additionally and more specifically, currently, Germany focuses on domestic burglary reports for its predictive policing tools. It is not clear if such reports reinforce the same cycle, as previously described. Studies that are more generally focused on reported domestic burglaries indicate that there is no significant difference in domestic burglary reports in poorer or richer communities (Hunter & Tseloni 2016; Oberwittler & Gerstner 2011: 47). Suffice to say, there is a significant lack of studies on this issue.

… Credit Scores

Policing in certain neighbourhoods already suggests the importance of residency as a racial (and social) marker. Hence, using residency as a marker in other AI prediction tools has the potential to, indirectly, reinforce ethnic disparities. Credit scores, either for individuals or for businesses, rely heavily on AI-driven risk assessment models. So, while most credit scoring companies, including the German credit scoring company SCHUFA, are not allowed to include indicators such as ethnicity, gender, income, or education, other factors such as residency can still indirectly discriminate against ethnic minorities (Campbell-Verduyn et al. 2017; Steinbauer 2005). Credit scores are an essential part of the lives of many, as these scores determine access to loans, the level of interest rates, as well as access to housing (Hinz & Auspurg 2017; Schlotböller 2020). Therefore, credit scores have always been scrutinised for potentially reaffirming existing prejudices and thus allowing discrimination to seep in without an adequate possibility of oversight, as the systems credit scores are based on are treated as trade secrets. As credit scoring systems are more and more driven by AI, concerns are raised that this may amplify such issues (Spader 2010).

While some studies on modern business credit scores find either no or only very little discrimination against ethnic minorities (Braunstein 2010; Robb & Robinson 2018), others report higher rejection rates among otherwise equivalent businesses owned by ethnic minorities compared to businesses owned by White people (Cavalluzzo & Wolken 2005; Henderson et al. 2015). For example, an analysis of credit scores for business start-ups in the US, based on the nationally representative Kauffman Firm Survey, revealed that Black and Latinx business owners had significantly

lower credit scores – even if other factors such as business-type, work experience, age, and more were accounted for (Henderson et al. 2015).

However, as these analyses cannot detangle how the opaque AI credit scoring systems come to discriminatory conclusions, it is hard to say what factors influence this undesirable outcome – even though residency is very likely part of it (Campbell-Verduyn et al. 2017). There is a legitimate concern that ethnic minorities might receive higher rates of false-positives for not being able to pay back loans (Henderson et al. 2015). This mirrors issues discussed in the former section on justice and rates of recidivism. Moreover, as statistical credit scoring systems have been around for centuries, the development of AI credit scoring systems seems to carry on discriminatory mechanisms observed for decades (Bates 1997).

... Healthcare

Racial biases are also deeply ingrained in healthcare systems: ethnic minorities and poorer communities have been under-researched and underfunded when it comes to medical treatments in hospitals in their prospective communities. While there is a reckoning in medicine overall, it comes as no surprise that the introduction of AI systems into healthcare and research has reinforced existing biases (Chase 2020; Ledford 2019; Rai 2019). AI is used for predictions of levels of hospital readmissions, mortality, and healthcare needs, as well as for image analysis, for example in radiology (Walsh et al. 2020).

In general, AI systems in healthcare seem to suffer primarily from narrow, unrepresentative data that either lacks participants from ethnic minorities or that does not report on the ethnicity of individuals. Additionally, the tendency to under-report symptoms for ethnic minorities due to misdiagnosis or social stigma impacts the quality of data available (Chase 2020; Walsh et al. 2020). This is a historically grown bias that is especially poignant in the US, which is where a majority of research on the topic has been conducted. However, there are analogies to be made for Germany, as it is not exempt from structural, institutional, and historical racial bias in healthcare (Wanger et al. 2020).

The concern about racially biased AI is not unfounded. For example, a commonly used AI system for prediction of healthcare needs in the US has been shown to be significantly biased against Black patients, as a study by Obermeyer et al. (2019) revealed. The issue, again, arose from scores given for individuals. Black people were disproportionately disadvantaged in this case because they were awarded too low of a score to receive the additional help they needed. The AI system used predicted health care costs as a basis for the level of illness/need. This has proven to be inherently flawed. While health care costs are usually a good predictor, historically, much less money is spent on sick Black people than sick White people. Hence, this prediction system reinforces the racial bias that is inherent the operation of the US system

(Obermeyer et al. 2019). However, Obermeyer et al. also provided a solution for this issue, as will be discussed in the chapter »Data Diversification and Transformation« on possible solutions.

… Job Markets

AI prediction systems are also deployed in hiring practices, specifically in the screening of potential candidates. In general, hiring practices have a long history of bias, especially against women and ethnic minorities (Bertrand & Mullainathan 2004; Raghavan et al. 2019), and recent studies show that not much has changed in the last 25 years (Quillian et al. 2017). Again, AI systems might reinforce discrimination that is inherent in historically discriminatory data (Cowgill 2018).[5]

The AI tools implemented during recruiting are mainly CV screening, automated online assessment, and automated screening of video-interviews (Lochner & Preuß 2018). As AI-based tools are usually not open to the public or researchers, it is quite often hard to tell if and how racial bias is inherent in AI hiring systems (Raghavan et al. 2019). But companies usually implement AI into their hiring processes with the aim of increasing fairness instead of perpetuating racial bias. Some companies have reported an increase in the diversity of their candidate pools after using AI (Houser 2019). However, these are self-reported increases.

Other studies have shown that racial bias is present in AI recruiting systems due to their heavy reliance on data from White candidates, especially for jobs that used to be held and are held primarily by White men (Ntoutsi et al. 2020). This mainly applies for CV and cover letter screenings, and less for automated online assessment. AI systems used for automated online assessment seem to be less susceptible to racial bias, as applicant data outside of the actual action performed does not come into play (Lochner & Preuß 2018; Ntoutsi et al. 2020). Additionally, recruiters, their bosses, and their boards often have a different definition of what ›fairness‹ entails in hiring practices and what constitutes discriminatory behaviour. It is difficult to implement AI systems to combat such issues when there is no coherent understanding of a fair selection process. This has been observed by ethnographers who closely followed the implementation of AI systems for recruiting in a big European firm (Sergeeva et al.

5 The underlying mechanisms are manifold and highly discussed. The two main economic models for labour discrimination are taste-based discrimination (Becker 1971) and statistical discrimination (Arrow 1971; Phelps 1972). The former assumes that there are emotional, irrational motives at play when discriminating against ethnic minorities. The latter assumes that discriminative employment decisions are the outcome of ›rational‹ actions based on the incomplete information that is available (Guryan/Charles 2013). Both interrogate how individuals make judgments as well as if and how humans can act ›rationally‹ – which is closely related to issues of human and AI bias.

2019). Again, AI systems seem to reveal existing issues of discrimination and racial bias.

... Consumer Goods and Miscellaneous

Additionally, there is an immense number of examples of racial bias in ›smaller‹ AI systems. For example, an algorithm trained on online human writing, associated ›pleasant‹ with White-sounding names and ›unpleasant‹ with Black-sounding names (Benjamin 2019). The search results of Google used to associate Black names with arrest records at much higher rates than White names (Benjamin 2019; Sweeney 2013). Another famous example is the facial recognition of Black faces; all popular facial recognition software is much less likely to correctly identify Brown and Black faces (Buolamwini & Gebru 2018). Another racial bias is to be seen in the voices used for AI systems. Nearly all voices can be classified as sounding ›White‹ which means they are accent-free and have a certain vocal performance that is usually associated with White people (Eidsheim 2018; Kushins 2014). As explored at the beginning of this paper, White voices are perceived as the most neutral setting and hence the most easily digestible voice (Caliskan et al. 2017). This list could go on and on but mostly reaffirms what has been stated at the beginning of this chapter. As it is hopefully apparent by now, racial bias can take hold in any technology developed and deployed by humans. As AI has the potential to amplify this bias, to obscure its origins and thereby ›neutralise‹ racial biases, there is certainly a need to propose solutions for how to move forward.

Proposed Solutions in Literature

Researchers from various fields have pursued different avenues in response to the question of how to address (racial) bias in AI. In the following, the most common recommendations will be addressed. These recommendations include: technological solutions, such as re-building algorithms and transforming/diversifying data; the regulation of AI in service of greater transparency; the education of researchers, developers, and practitioners; the diversification of research and development teams.

Re-building Algorithms

As previously established, a distinction can be made for algorithmic and data bias. For both, there are proposed technological solutions. One option to combat algorithmic bias is to ›re-build‹ algorithms in a way that shows fewer (racial) bias. Instances of adapted regression models and technological frameworks have been given, for example, by Bower et al. (2018) and Kleinberg et al. (2018). In general, in order to re-

build algorithms in a way that minimises racial bias, a sufficient understanding of how and why biases occur in the first place is necessary. Here, issues of transparency come into play. While sometimes the algorithms underlying AI systems themselves are the culprits, more often than not the main issue seems to lie in the data fed into AI systems.

Data Diversification and Transformation

When it comes to data bias,[6] some solutions have already been tested and put into place. The main approach is to diversify the data sets used for training AI by bringing these sets into accord with the ›real‹ population or using data sets that are more inclusive. Several computer scientists have developed new data sets for different fields. One example is to diversify data sets to train common face recognition software, as proposed by Buolamwini and Gebru (2018). Another approach for AI systems in healthcare, as already alluded to, was tried and tested by Obermeyer et al. (2019): together with the developers of the scrutinised algorithm, they transformed the data fed to the algorithm. Instead of relying on predicted future cost (as was explored earlier), they created a variable that combined predicted cost with predicted health. This little adjustment decreased the observed bias by 84 %. Here, it was not a simple case of diversification through the addition of more people of colour into the database but rather a re-adjustment of the labels given. Such tangible and successful scientific interventions in AI systems by private entities are rare though, or at least they are rarely reported in scientific journals.

Another option is to transform data sets into synthetic data sets. In short, this means that data of real individuals (for example in healthcare) is gathered and transformed into a set of new data points with the same overall properties before it is fed into (primarily deep learning) algorithms. The synthetic data set usually mirrors the statistical properties of the original data and preserves the original data structure. This technique is primarily used to increase data privacy and allows one to build software without exposing personal data to developers, as they only work with the synthetic set. Additionally, first tests show that synthetic data sets might balance biased data sets (Gretel.AI & Watson 2020; Kortylewski et al. 2019). Adding synthetically generated patient records to training sets for health care interventions, for example, increased the accuracy of classification algorithms in most tested algorithms. In this test, gender bias was recorded and reduced after adding »synthetic patients« (Gretel.AI & Watson 2020). However, there is certainly a lack of scientific papers on whether and how synthetic data might be useful for reducing racial bias.

6 At this point, it should be noted that ›bias‹ can have a very different meaning in computer science literature, as it refers to technological bias, not to racial/algorithmic data biases as defined in this expertise.

Additionally, while technological solutions are necessary, they do not just ›happen‹. A successful technological intervention is only possible when programmers either realise the bias inherent in their algorithms or data and/or have the incentive to change the outcome. One cannot always rely on researchers finding bias in AI systems and on being welcomed by the original developers when one attempts to find solutions, as happened in the case of Obermeyer et al. (2019). Hence, there is a myriad of proposed solutions that revolve more around policy solutions and different types of intervention in the development process in order to allow for a more thorough and holistic approach to detecting and reducing (racial) bias in AI.

Regulation and Transparency

One solution to combat biases and to allow for a better analysis is a more comprehensive and focused regulation of AI. Several non-profit organisations, legal and policy scholars, and advisory boards have put forward suggestions for the regulation of AI.

The European Commission put forward its AI White Paper in 2020 and followed up with a public consultation that included stakeholders from civil society, industry, and academia. The in-depth results will be put forward in the first quarter of 2021. So far, regulatory efforts are planned but not fully sketched out (European Commission 2021). In general, the commission states that AI is to be held to the same standards as all technology. The paper highlights one of the main issues of successful regulation of AI: verification of compliance with EU law, since AI systems are often opaque, complex, and unpredictable to a certain extent (European Commission 2020: 12).

Others have put forward more concrete demands for regulation. All these demands have in common that their main concern is how to increase the transparency of AI and how to provide access to data. Recurring themes that are relevant to combatting racial bias are (AlgorithmWatch & Bertelsmann Stiftung 2020; Erdélyi & Goldsmith 2018; Iphofen & Kritikos 2019; Kemper & Kolkman 2019; Lupo 2019; Raghavan et al. 2019; Scherer 2015; Vincent & Viljoen 2020):

- **Public registers for AI systems:** The public sector has a higher need for legitimacy and accountability. Hence, there are calls for a public register of AI systems that are used in public administration; this register may include information on if and where AI is used, what kind of AI is used, what data sets are included, and who developed it. Furthermore, many call for a similar level of disclosure of AI systems deployed by private entities if these entities have a significant impact on humans (for example credit scoring).
- **Data access frameworks:** Another regulatory proposal is to allow for access to training data and data results. Again, this is meant to help the public and scientific understanding. As highlighted before, there is a lack of empirical research on how current AI systems function and with what data they were

trained. Hence, it is often hard to grasp if and how racial bias is inherent in AI systems.

- **Audits of AI systems:** Many call for an auditing process that systematically and independently screens and documents AI systems, especially concerning transparency of and discrimination through AI. What exactly this entails and how this independent entity should be built and managed is usually not defined in full detail. Most scholars and civil society organisations argue for an inclusive stakeholder process to establish a framework for auditing systems.
- **Humans are in the loop:** One regulatory demand is that decisions always have to stay in the hands of humans. Hence, AI concepts where humans would be out of the loop, as explained in the chapter »Definition of AI«, should generally be forbidden in the eyes of most researchers.
- **Inclusion of civil society and academia**: Overall, most researchers and civil society organisations demand the inclusion of a critical audience in all regulatory processes and auditing frameworks to allow for meaningful enforcement and monitoring of transparent AI.

However, there are limits to the call for transparency. Especially in academic discourse, there is a certain understanding that full transparency is neither possible nor feasible. To just supply the outline of AI systems, and maybe even their code, is often not enough to understand the decisions made by AI and how it arrived at them (Chiao 2019). Some AI systems, especially deep neural nets, are so opaque that not even their developers fully understand their output. Deep neural nets use billions of parameters that are nested and interconnected, hence, any linear connection between input and output gets muddied: One can see that input (a) led to output (a) but not fully understand how and why the neural net arrived at this conclusion. To understand how neural nets function one can use this online tool. Here, some call for explainable AI, which is based on the idea that every AI system needs to provide reasoning for how it arrived at its results. However, questions remain about how detailed and how accessible (in the sense of comprehensiveness for either experts, interested parties, or laypeople) such explanations should and would be (Zarsky 2016). Many argue that a radical approach to fairness might not be feasible and that such an approach asks more of AI than of any other form of decision making. In a sense, every decision is opaque to some extent, as no one can look inside the mind of a human judge, for example, even though a judge provides written reasoning. While humans can defend their decisions by reasoning, the reasoning of AI systems might follow a logic that humans cannot fully grasp (Lee & Björklund Larsen 2019; Zarsky 2016).

Education of Practitioners, Researchers, and Developers

Another more straightforward approach to combatting racial bias in AI is to provide training. On the one hand, scholars ask that practitioners who interact with AI system, especially in legal settings, receive education in order to understand the basic functionality of applied AI systems and to gain awareness of potential (racial) bias (Hart 2016; Vincent & Viljoen 2020). On the other hand, AI researchers and developers need to gain a basic understanding of the societal impact their inventions might have. In both cases, scholars tend to argue for the inclusion of such topics in the curricula of associated study courses as well as for (mandatory) further education programmes on the job (AlgorithmWatch & Bertelsmann Stiftung 2020; Iphofen & Kritikos 2019; Lupo 2019).

Diversification of AI Research and Development

Closely connected to better education is the realisation that most programming teams are still quite homogenous – very often White, even more often male. Hence, the people developing AI are usually from a similar cultural and social background (West et al. 2019). However, a lack of diversity can lead to several problems. Firstly, it might reduce the drive for innovation. Additionally, a lack of diversity might lead to a lack of understanding and consideration for people outside one's bubble (Levine et al. 2014). This has been shown to be one of the reasons why certain AI systems either have not been trained to handle data of non-Whites or have not been trained with data of non-Whites (Bacchini & Lorusso 2019; Buolamwini & Gebru 2018). Teams that are not ›diverse‹ tend to re-produce existing (power) imbalances (Freire et al. 2020; Houser 2019). Hence, one of the proposed solutions is to foster more diverse working environments. In a team with different (life) experiences and sensibilities, more and diverging concerns are taken into account. Simply put, if you have a Black person in your research team, it is more likely that a flag will be raised if you train your algorithm with primarily White pictures.

However, there are two caveats to this assumption. First, a diverse team does not necessarily lead to a diverse mindset – or (racial) issues might be spotted but then disregarded when the social setting is considered. For example, a Black programmer reported that he was aware of the cliché/implications of using White-sounding voices for AI service assistants. However, he did not object because he ›knew‹ that proposing to use a Black-sounding voice would be seen as a ›political‹ act and might reduce a buyer's interests (Benjamin 2019). As part of an underrepresented minority, one might be more aware of bias and problems – however, as a human still acting inside a society with structural racism, one is not exempt from harvesting unconscious bias. As a person who has to think in economical ways, one might still reproduce racist stereotypes or reinforce racial biases in decision-making processes.

Additionally, the burden of speaking-up does not lie on the individual in the perceived minority group. The idea of getting rid of racial bias by diversifying a team is flawed as it neglects the responsibility of White people to speak up when bias creeps in. A diverse team is an important goal in itself, not a way to get oneself out of difficult thought-processes and conversations. Second, questions remain about how to enforce a more diverse team. Some call for fixed quotas, while many others see a more feasible approach in raising awareness and providing incentives for a more diverse workforce (Houser 2019; Leggon 2010; West et al. 2019). However, this might be a very tedious and lengthy process, the effectiveness of which will need constant reevaluation.

Overall, all proposed solutions to combat (racial) bias are not exclusive but rather are most likely to succeed if they are interwoven and simultaneously pursued.

AI as Part of the Solution

Despite the many issues highlighted up to now, AI might still have the potential to reduce biases rather than amplify or reproduce them (Cowgill 2018; Houser 2019). As it is the system and/or humans that are racist, not AI, careful and conscious deployment of AI systems might be able to actively mitigate racial bias (Vincent & Viljoen 2020). Therefore, some researchers argue that the focus on problems with AI negates the potential of AI. They ask for a certain shift of perspective when thinking about racism and AI: rather than seeing AI as the culprit, one might be better advised to just see it as the technological tool it is. Positive examples are to be found in hiring practices, as alluded to before (Cowgill 2018; Raghavan et al. 2019). Additionally, AI might also be an active tool to identify and combat acts of racism, for example, hate speech and racist harassment online (Benjamin 2019: 13–14).

Conclusion and Research Needs

Overall, AI often acts as an amplifier of existing injustices. Furthermore, it tends to make racial bias more visible even though this bias was always there. AI, and technology in general, are shaped by humans and the environment they are placed in. To understand racism and AI, one has to understand human racial bias and racist beliefs, structures, and institutions. Still, AI requires us to face some unique challenges and questions. While the research on AI and racism is growing, there is still a lot to be discovered, analysed, and understood – especially in Germany. Therefore, the following research recommendations are given:

- **More research in Germany and the European Union:** Most research that directly targets racism and AI comes from the US. While there are similarities and analogies to be made for Germany and the European Union, there are different and fewer AI systems in place in these areas (especially in policing/justice context) than in the US. Additionally, structural, institutional, individual racism presents differently in the European Union overall and Germany specifically. Lastly, there are other legal requirements for technology development, which might shift the questions that need to be answered. While there are German scholars who tackle racial bias in AI (e.g., Hagendorff, 2019), there is a huge lack of research on this intersection, especially empirical research.
- **More case studies:** While there is a lot of speculation about and many singular examples of how and why racial bias appears in AI systems, it might be of interest to work more closely with AI developers. AI systems (developed and in use in Germany and Europe) should be analysed in a more structured way, with an emphasis on potential racial biases.
- **More interdisciplinary research:** To successfully understand and potentially combat racial bias, closer exchange and joint research are needed. We still need to fully grasp exactly how AI contributes to racial bias – if it just amplifies, reinforces, or adds a new layer. As seen before, most issues of racial bias in AI intersect with several research fields and ask questions about historical and on-going human bias as well as technological issues that are inherent in AI. Hence, all covered areas would profit from more interdisciplinary research on how human bias influences AI, and vice-versa, and what possible solutions might look like. Additionally, the inherent issue of the explainability of certain AI (black box) requires a deeper exploration of how we define accountability in AI and if and how explainability can be technically provided. This can only be successfully addressed by including researchers from all affected fields, such as law, public policy, sociology, philosophy, and computer science.
- **More intersectional research:** AI potentially amplifies any human bias. Hence, it is advisable to ›scan‹ AI systems for biases towards ethnic minorities, people with disabilities, older people, poorer people, specific religious groups, genders and sexualities deviating from heteronormative assumptions, and many more. As one person can hold several social and political identities, different modes of discrimination might combine. To fully understand how this is at play when it comes to bias in AI, intersectional approaches should be taken (for example, ethnicity and class often interplay when it comes to the policing of certain communities; being female and Black might lead to specific discrimination in AI healthcare systems).
- **More research on predictive policing and racism:** This, again, is especially a recommendation for Germany – and the European Union for that matter. The pre-

dictive policing technologies already deployed should be scientifically analysed with a special focus on potential racial bias and other discriminatory factors.

- **More research on AI systems in healthcare and racism:** As with the former point, this is especially a research need in Germany and the European Union. A better understanding is needed of how the AI systems in place in healthcare might mitigate and deepen racial bias that is present in medicine and healthcare.
- **More research on racism and smart cities:** So far, the intersection of racism and city planning mainly revolves around issues of policing and potential redlining. We need a more thorough understanding of how AI systems that shape (future) cities might reinforce and/or amplify racial discrimination and disparities. First theoretical explorations have been conducted (e.g., Del Casino 2016; Safransky 2020) but there is a need for empirical data and case studies on this issue in particular, especially in urban planning and research on smart cities.
- **More research on the diversification of research teams:** As highlighted before, there is already research on the lack of diversity in research and development teams in the fields encompassing AI research. But there are few experiments on or accompanying observations of the effects of more diverse teams, especially considering a possible reduction/stronger consideration of racial bias.
- **A deeper exploration of ex-ante regulation of AI:** In the current scholarly discourse on how to build ›fairer‹ AI, a lot of focus is set on ex-post (after the development of AI) regulation solutions. It might be of use to also focus on ex-ante (before the development) approaches to ›fairer‹ AI from a scientific standpoint. This could take place, for example, through the establishment of ethical AI guidelines, as has been put forward by some non-profit organisations (e.g., Bertelsmann Stiftung 2020). Again, more interdisciplinary research might strengthen this point.

References

Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., Hall, J. S., Samsonovich, A., Scheutz, M., Schlesinger, M., Shapiro, S. C., & Sowa, J. (2012): »Mapping the Landscape of Human-Level Artificial General Intelligence.« In: AI Magazine 33, 25–42. https://doi.org/10.1609/aimag.v33i1.2322

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019): »Exploring the Impact of Artificial Intelligence: Prediction versus Judgment.« In: Information Economics and Policy, The Economics of Artificial Intelligence and Machine Learning 47, 1–6. https://doi.org/10.1016/j.infoecopol.2019.05.001

AlgorithmWatch & Bertelsmann Stiftung (eds.) (2020): »Automating Society Report 2020 – Deutschland.« https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/automating-society-report-2020-deutschland

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016): How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm?token=SV45W9VHgigYbUE-m709xnvExqobnjcg

Arnold, D., Dobbie, W., & Hull, P. (2020): »Measuring Racial Discrimination in Bail Decisions.« Becker Friedman Institute, Working Paper 2020–33.

Arrow, K. (1971): Some Models of Racial Discrimination in the Labor Market. Rand.

Bacchini, F., & Lorusso, L. (2019): »Race, Again: How Face Recognition Technology Reinforces Racial Discrimination.« In: Journal for Information, Communication & Ethics in Society 17, 321–335. https://doi.org/10.1108/JICES-05-2018-0050

Bates, T. (1997): »Unequal Access: Financial Institution Lending to Black-and White-Owned Small Business Start-Ups.« In: Journal of Urban Affairs 19, 487–495. https://doi.org/10.1111/j.1467-9906.1997.tb00508.x

Beck, S. (2020): »Künstliche Intelligenz – ethische und rechtliche Herausforderungen.« In: K. Mainzer, (ed.), Philosophisches Handbuch Künstliche Intelligenz. Springer Fachmedien, 1–28. https://doi.org/10.1007/978-3-658-23715-8_29-1

Becker, G. S. (1971): The Economics of Discrimination (2nd ed.). Chicago: University of Chicago Press.

Benjamin, R. (2019): Race after Technology: Abolitionist Tools for the New Jim Code. Newark, UK: Polity Press.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018): »Fairness in Criminal Justice Risk Assessments: The State of the Art.« In: Sociological Methods & Research 50(1), 3–44. https://doi.org/10.1177/0049124118782533

Bertelsmann Stiftung (2020): Regeln für die Gestaltung algorithmischer Systeme. *Algo.Rules*. https://algorules.org/de/startseite

Bertrand, M., & Mullainathan, S. (2004): »Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.« In: Amercian Economic Review 94, 991–1013. https://doi.org/10.1257/0002828042002561

Bower, A., Niss, L., Sun, Y., & Vargo, A. (2018): »Debiasing Representations by Removing Unwanted Variation Due to Protected Attributes.« ArXiv180700461 Cs.

Braunstein, S. (2010): »Credit Scoring: Testimony before the Subcommittee on Financial Institutions and ConsumerCredit, the Committee on Financial Services.« In: Testimony before the Subcommittee on Financial Institutions and Consumer Credit Financial Services Committee, U.S. House of Representatives.

Buolamwini, J., & Gebru, T. (2018): »Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.« In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017): »Semantics Derived Automatically from Language Corpora Contain human-like Biases.« In: Science 356, 183–186. https://doi.org/10.1126/science.aal4230

Campbell-Verduyn, M., Goguen, M., & Porter, T. (2017): »Big Data and Algorithmic Governance: The Case of Financial Practices.« In: New Political Economy 22, 219–236. https://doi.org/10.1080/13563467.2016.1216533

Cavalluzzo, K., & Wolken, J. (2005): »Small Business Loan Turndowns, Personal Wealth, and Discrimination.« In: The Jounral of Bussiness 78, 2153–2178. https://doi.org/10.1086/497045

Cave, S., & Dihal, K. (2020): »The Whiteness of AI.« In: Philosophy & Technology 33, 685–703. https://doi.org/10.1007/s13347-020-00415-6

Chase, A .C. (2020): »Ethics of AI: Perpetuating Racial Inequalities in Healthcare Delivery and Patient Outcomes.« In: Voices in Bioethics 6, https://doi.org/10.7916/vib.v6i.5890

Chiao, V. (2019): »Fairness, Accountability and Transparency: Notes on Algorithmic Decision-making in Criminal Justice.« In: International Journal of Law in Context 15, 126–139. https://doi.org/10.1017/S1744552319000077

Cowgill, B. (2018): »Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening.« In: Columbia Business School, Columbia University, 2018, 29. https://conference.iza.org/conference_files/MacroEcon_2017/cowgill_b8981.pdf

Danaher, J. (2016): »The Threat of Algocracy: Reality, Resistance and Accommodation.« In: Philosophy & Technology 29, 245–268. https://doi.org/10.1007/s13347-015-0211-1

Del Casino, V. J. (2016): »Social Geographies II: Robots.« In: Progress in Human Geography 40, 846–855. https://doi.org/10.1177/0309132515618807

Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L. C., Misra, S., Mogaji, E., Sharma, S. K., Singh, J. B., Raghavan, V., Raman, R., Rana, N. P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., & Williams, M. D. (2019): »Artificial Intelligence (AI): Multidisciplinary Perspectives on Emerging Challenges, Opportunities, and Agenda for Research, Practice and Policy.« In: International Journal of Information Management 57, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Egbert, S., & Krasmann, S. (2020): »Predictive Policing: Not yet, but Soon Preemptive?« In: Policing and Society 30, 905–919. https://doi.org/10.1080/10439463.2019.1611821

Eidsheim, N. S. (2018): The race of sound: listening, timbre, and vocality in African American music, Refiguring American music. Durham, NC: Duke University Press.

Erdélyi, O. J., & Goldsmith, J. (2018): »Regulating Artificial Intelligence: Proposal for a Global Solution.« In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES ›18. Association for Computing Machinery, 95–101. https://doi.org/10.1145/3278721.3278731

European Commission (2021): »Shaping Europe's Digital Future.« https://ec.europa.eu/digital-single-market/en/artificial-intelligence

European Commission (2020): »White paper on Artificial Intelligence: A European Approach to Excellence and Trust.« https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Franklin, S. (2014): »History, Motivations, and Core Themes.« In: K. Frankish, W. M. Ramsey (eds.), The Cambridge Handbook of Artificial Intelligence. Cambridge University Press, 15–33. https://doi.org/10.1017/CBO9781139046855.003

Freire, A., Porcaro, L., & Gómez, E. (2020): »Measuring Diversity of Artificial Intelligence Conferences.« ArXiv200107038 Cs.

Gretel.AI, & Watson, A., (2020): Reducing AI bias with Synthetic data. *Gretel.AI*. https://gretel.ai/blog/reducing-ai-bias-with-synthetic-data

Guryan, J., & Charles, K. K. (2013): »Taste-Based or Statistical Discrimination: The Economics of Discrimination Returns to its Roots.« In: The Economic Journal 123, F417–F432. https://doi.org/10.1111/ecoj.12080

Hagendorff, T. (2019): »Rassistische Maschinen?« In: M. Rath, F. Krotz, & M. Karmasin (eds.), Maschinenethik: Normative Grenzen autonomer Systeme, Ethik in mediatisierten Welten. Springer Fachmedien, 121–134. https://doi.org/10.1007/978-3-658-21083-0_8

Hao, K., & Stray, J. (2019): »Can You Make AI Fairer Than a Judge? Play our Courtroom Algorithm Game.« In: MIT Technology Review. https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

Hart, S. D. (2016): »Culture and Violence Risk Assessment: The Case of Ewert v. Canada.« In: Journal of Threat Assessment and Management 3, 76–96. https://doi.org/10.1037/tam0000068

Henderson, L., Herring, C., Horton, H. D., & Thomas, M. (2015): »Credit Where Credit is Due?: Race, Gender, and Discrimination in the Credit Scores of Business Startups.« In: The Review of Black Political Economy 42, 459–479. https://doi.org/10.1007/s12114-015-9215-4

Hinz, T., & Auspurg, K. (2017): »Diskriminierung auf dem Wohnungsmarkt.« In: A. Scherr, A. El-Mafaalani, G. Yüksel (eds.), Handbuch Diskriminierung. Springer Fachmedien, 387–406. https://doi.org/10.1007/978-3-658-10976-9_21

Houser, K. A. (2019): »Can AI Solve the Diversity Problem in the Tech Industry? Mitigating Noise and Bias in Employment Decision-Making.« Stanford Technology Law Review 22, 290–353. https://ssrn.com/abstract=3344751

Hunter, J., & Tseloni, A. (2016): »Equity, Justice and the Crime Drop: The case of Burglary in England and Wales.« In: Crime Science 5, 3. https://doi.org/10.1186/s40163-016-0051-z

Iphofen, R., & Kritikos, M. (2019): »Regulating Artificial Intelligence and Robotics: Ethics by Design in a Digital Society.« In: Contemporary Social Science 16(2), 1–15. https://doi.org/10.1080/21582041.2018.1563803

Jefferson, B. J. (2018): »Policing, Data, and Power-geometry: Intersections of Crime Analytics and Race during Urban Restructuring.« In: Urban Geography 39, 1247–1264. https://doi.org/10.1080/02723638.2018.1446587

Jiang, S. (2020): »Automatisierte Entscheidungsfindung, Strafjustiz und Regulierung von Algorithmen. Ein Kommentar zum Fall ›State v. Loomis‹.« In: Beck, S., Kusche, C., Valerius, B.: Digitalisierung, Automatisierung, KI und Recht. Nomos Verlagsgesellschaft mbH & Co. KG, 557–590. https://doi.org/10.5771/9783748920984-557

Jobin, A. (2017): »Von A(pfelkuchen) bis Z(ollkontrolle): Weshalb Algorithmen nicht neutral sind (From A(pple pie) to Z(ealous Customs Control): Why Algorithms Are Not Neutral).« In: A. Fichter (ed.), Smartphone-Demokratie. NZZ Libro, 154–170. https://doi.org/10.5281/zenodo.1198590

Jones, L. (2020): »A Philosophical Analysis of AI and Racism.« In: Stance 13, 36–46. https://doi.org/10.5840/stance2020133

Katyal, S. K. (2019): »Private Accountability in the Age of Artificial Intelligence.« In: UCLA Law Review 66, 54.

Kearns, M., & Roth, A. (2019): The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford: Oxford University Press.

Kemper, J., & Kolkman, D. (2019): »Transparent to Whom? No Algorithmic Accountability without a Critical Audience.« In: Information, Communication & Society 22, 2081–2096. https://doi.org/10.1080/1369118X.2018.1477967

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018): »Algorithmic Fairness.« In: AEA Papers and Proceedings 108, 22–27. https://doi.org/10.1257/pandp.20181018

Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., & Vetter, T. (2019): »Analyzing and Reducing the Damage of Dataset Bias to Face Recognition With Synthetic Data.« 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2261–2268.Kushins, E. R. (2014): »Sounding Like Your Race in the Employment Process: An Experiment on Speaker Voice, Race Identification, and Stereotyping.« In: Race and Social Problems 6, 237–248. https://doi.org/10.1007/s12552-014-9123-4

Ledford, H. (2019): »Millions of Black People Affected by Racial Bias in Health-care Algorithms.« In: Nature 574, 608–609. https://doi.org/10.1038/d41586-019-03228-6

Lee, F., & Björklund Larsen, L. (2019): »How Should We Theorize Algorithms? Five Ideal Types in Analyzing Algorithmic Normativities.« In: Big Data & Society 6, 2053951719867349. https://doi.org/10.1177/2053951719867349

Leggon, C. B. (2010): »Diversifying Science and Engineering Faculties: Intersections of Race, Ethnicity, and Gender.« In: American Behavioral Scientist 53, 1013–1028. https://doi.org/10.1177/0002764209356236

Levine, S. S., Apfelbaum, E. P., Bernard, M., Bartelt, V .L., Zajac, E. J., & Stark, D. (2014): »Ethnic Diversity Deflates Price Bubbles.« In: Proceedings of the National Academy of Sciences of the United States of America, 111(52), 18524–18529.

Lochner, K., & Preuß, A. (2018): »Digitales Recruiting.« In: Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO) 49, 193–202. https://doi.org/10.1007/s11612-018-0425-7

Lupo, G. (2019): »Regulating (Artificial) Intelligence in Justice: How Normative Frameworks Protect Citizens from the Risks Related to AI Use in the Judiciary.« In: European Quarterly of Political Attitudes and Mentalities, 8 (2), 75–96. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-62463-8

MacKenzie, D. A., & Wajcman, J. (eds.) (1999): The social shaping of technology (2nd ed.). Buckingham, UK; Philadelphia, PA: Open University Press.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., & Staab, S. (2020): »Bias in Data-driven Artificial Intelligence Systems – An Introductory Survey.« In: WIREs Data Mining and Knowledge Discovery 10, e1356. https://doi.org/10.1002/widm.1356

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019): »Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.« In: Science 366, 447–453. https://doi.org/10.1126/science.aax2342

Oberwittler, D., & Gerstner, D. (2011): Kriminalgeographie Baden-Württembergs (2003–2007): sozioökonomische und räumliche Determinanten der registrierten Kriminalität. Freiburg i.Br.: Max-Planck-Inst. für Ausländisches und Internat. Strafrecht.

Pasquale, F. (2015): The Black Box Society: The Secret Algorithms That Control Money and Information (1st ed.). Cambridge, MA: Harvard University Press.

Phelps, E. (1972): »The Statistical Theory of Racism and Sexism.« In: American Economic Review 62, 659–61.

Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. (2017): »Meta-analysis of Field Experiments Shows No change in Racial Discrimination in Hiring over Time.«

In: Proceedings of the National Academy of Sciences of the United States of America, 114(41), 10870–10875. https://doi.org/10.1073/pnas.1706255114

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2019): »Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices.« ArXiv190609208 Cs. https://doi.org/10.1145/3351095.3372828

Rai, T. S. (2019): »Racial Bias in Health Algorithms.« In: Science 366, 440–441. https://doi.org/10.1126/science.366.6464.440-e

Robb, A., & Robinson, D. T. (2018): »Testing for Racial Bias in Business Credit Scores.« In: Small Business Economics 50, 429–443. https://doi.org/10.1007/s11187-017-9878-2

Safransky, S. (2020): »Geographies of Algorithmic Violence: Redlining the Smart City.« In: International Journal of Urban and Regional Research 44, 200–218. https://doi.org/10.1111/1468-2427.12833

Scherer, M. U. (2015): »Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies.« In: Harvard Journal of Law & Technology 29, 353.

Schlotböller, D. (2020): »Diskriminierung im Sozialkreditsystem.« In: O. Everling (ed.), Social Credit Rating: Reputation und Vertrauen beurteilen. Springer Fachmedien, 331–345. https://doi.org/10.1007/978-3-658-29653-7_18

Seidensticker, K., Bode, F., & Stoffel, F. (2018): »Predictive Policing in Germany.« In: Konstanzer Online-Publikations-System. https://kops.uni-konstanz.de/entities/publication/4a808e3e-e518-47ec-807d-ca76e46e2a97

Sergeeva, A., Huysman, M., & van den Broek, E. (2019): »Hiring Algorithms: An Ethnography of Fairness in Practice.« In: International Conference on Information Systems (ICIS) 2019 Proceedings. The Future of Work 6 https://aisel.aisnet.org/icis2019/future_of_work/future_work/6

Shapiro, A. (2017): »Reform Predictive Policing.« In: Nature 541, 458–460. https://doi.org/10.1038/541458a

Skeem, J. L., & Lowenkamp, C. T. (2016): »Risk, Race, and Recidivism: Predictive Bias and Disparate Impact« In: Criminology 54, 680–712. https://doi.org/10.1111/1745-9125.12123

Sommerer, L. M. (2017): »Geospatial Predictive Policing – Research Outlook & A Call For Legal Debate.« In: Neue Kriminalpolitik 29, 147–164.

Spader, J. (2010): »Beyond Disparate Impact: Risk-based Pricing and Disparity in Consumer Credit History Scores.« In: The Review of Black Political Economy 37, 61–78. https://doi.org/10.1007/s12114-009-9055-1

Steinbauer, D. (2005): »An Intermediate Information System Forms Mutual Trust.« In: Härder, T., Lehner, W. (eds), Data Management in a Connected World. Lecture Notes in Computer Science, vol 3551. Springer, Berlin, Heidelberg, 277–292. https://doi.org/10.1007/11499923_15

Straube, T., & Belina, B. (2018): »Policing the Smart City: Eine Taxonomie polizeilicher Prognoseprogramme.« In: T. Straube & B. Belina (eds.), Smart City

– Kritische Perspektiven auf die Digitalisierung in Städten. transcript Verlag. https://doi.org/10.14361/9783839443361-016

Sweeney, L. (2013): »Discrimination in Online Ad Delivery.« https://arxiv.org/abs/1301.6822v1

Vincent, G. M., & Viljoen, J. L. (2020): »Racist Algorithms or Systemic Problems? Risk Assessments and Racial Disparities.« In: Criminal Justice and Behavior 47, 1576–1584. https://doi.org/10.1177/0093854820954501

Walsh, C. G., Chaudhry, B., Dua, P., Goodman, K. W., Kaplan, B., Kavuluru, R., Solomonides, A., & Subbian, V. (2020): »Stigma, Biomarkers, and Algorithmic Bias: Recommendations for Precision Behavioral Health with Artificial Intelligence.« In: JAMIA Open 3, 9–15. https://doi.org/10.1093/jamiaopen/ooz054

Wanger, L., Kilgenstein, H., & Poppel, J. (2020): »Über Rassismus in der Medizin.« In: Kritische Medizin München.

West, S. M., Whittaker, M., & Crawford, K. (2019): »Discriminating Systems: Gender, Race, and Power in AI.« AI Now Institute. https://ainowinstitute.org/discriminatingsystems.html

Zarsky, T. (2016): »The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making.« In: Science, Technology & Human Values 41, 118–132. https://doi.org/10.1177/0162243915605575