Chapter 2
Unfair and Illegal Discrimination

Abstract There is much debate about the ways in which artificial intelligence (AI) systems can include and perpetuate biases and lead to unfair and often illegal discrimination against individuals on the basis of protected characteristics, such as age, race, gender and disability. This chapter describes three cases of such discrimination. It starts with an account of the use of AI in hiring decisions that led to discrimination based on gender. The second case explores the way in which AI can lead to discrimination when applied in law enforcement. The final example looks at implications of bias in the detection of skin colour. The chapter then discusses why these cases are considered to be ethical issues and how this ethics debate relates to well-established legislation around discrimination. The chapter proposes two ways of raising awareness of possible discriminatory characteristics of AI systems and ways of dealing with them: AI impact assessments and ethics by design.

Keywords Discrimination · Bias · Gender · Race · Classification · Law enforcement · Predictive policing · AI impact assessment · Ethics by design

2.1 Introduction

Concern at discrimination is probably the most widely discussed and recognised ethical issue linked to artificial intelligence (AI) (Access Now 2018; Latonero 2018; Muller 2020). In many cases an AI system analyses existing data which was collected for purposes other than the ones that the AI system is pursuing and therefore typically does so without paying attention to properties of the data that may facilitate unfair discrimination when used by the AI system. Analysis of the data using AI reveals underlying patterns that are then embedded in the AI model used for decision-making. In these cases, which include our examples of gender bias in staff recruitment and predictive policing that disadvantages segments of the population, the system perpetuates existing biases and reproduces prior practices of discrimination.

In some cases, discrimination occurs through other mechanisms, for example when a system is exposed to real-world data that is fundamentally different from the data it was trained on and cannot process the data correctly. Our case of systems that

B. C. Stahl et al., *Ethics of Artificial Intelligence*,
SpringerBriefs in Research and Innovation Governance,
https://doi.org/10.1007/978-3-031-17040-9_2

misclassify people from ethnic groups that are not part of the training data falls into this category. In this case the system works in a way that is technically correct, but the outputs are incorrect, due to a lack of correspondence between the AI model and the input data.

These examples of AI-enabled discrimination have in common that they violate a human right (see box) that individuals should not be discriminated against. That is why these systems deserve attention and are the subject of this chapter.

Universal Declaration of Human Rights, Article 7

"All are equal before the law and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination." (UN 1948)

2.2 Cases of AI-Enabled Discrimination

2.2.1 Case 1: Gender Bias in Recruitment Tools

Recruiting new members of staff is an important task for an organisation, given that human resources are often considered the most valuable assets a company can have. At the same time, recruitment can be time- and resource-intensive. It requires organisations to scrutinise job applications and CVs, which are often non-standardised, complex documents, and to make decisions on shortlisting and appointments on the basis of this data. It is therefore not surprising that recruitment was an early candidate for automation by machine learning. One of the most high-profile examples of AI use for recruitment is an endeavour by Amazon to automate the candidate selection process.

In 2014, Amazon started to develop and use AI programs to mechanise highly time-intensive human resources (HR) work, namely the shortlisting of applicants for jobs. Amazon "literally wanted it to be an engine where I'm going to give you 100 résumés, it will spit out the top five, and we'll hire those" (Reuters 2018). The AI tool was trained on CVs submitted over an earlier ten-year period and the related staff appointments. Following this training, the AI tool discarded the applications of female applicants, even where no direct references to applicants' gender were provided. Given the predominance of successful male applicants in the training sample, Amazon found that the system penalised language such as "women's chess club captain" for not matching closely enough the successful male job applicants of the past. While developers tried to modify the system to avoid gender bias, Amazon abandoned its use in the recruitment process in 2015 as a company "committed to workplace diversity and equality" (ibid).

At first this approach seemed promising, as HR departments have ample training data in the form of past applications. A machine learning system can thus be trained to distinguish between successful and unsuccessful past applications and identify features of applications that are predictors of success. This is exactly what Amazon did. The result was that the AI systematically discriminated against women.

When it became clear that women were being disadvantaged by recruitment based on AI, ways were sought to fix the problem. The presumptive reason for the outcome was that there were few women in the training sample, maybe because the tech sector is traditionally male dominated, or maybe reflecting biases in the recruitment system overall. It turned out, however, that even removing direct identifiers of sex and gender did not level the playing field, as the AI found proxy variables that still pointed to gender, such as place of study (e.g., all-female college) and feminised hobbies.

> AI systems are only as good as the data they're trained on and the humans that build them. If a résumé-screening machine-learning tool is trained on historical data, such as résumés collected from a company's previously hired candidates, the system will inherit both the conscious and unconscious preferences of the hiring managers who made those selections (Heilweil 2019).

In the case of Amazon this eventually led to the company's abandoning the use of AI for hiring, as explained in the case description. However, the fundamental challenge of matching large numbers of candidates for recruitment with large numbers of open positions on the basis of complex and changing selection criteria remains. For instance, Vodafone is reported to have used AI systems to analyse over 100,000 graduate applications for 1,000 jobs (Kaur 2021). Since 2019, the COVID-19 pandemic has accelerated the use of AI recruitment, with predictions that 16% of HR recruitment jobs will have disappeared by 2029 (ibid).

AI can also, it is claimed, be used as a tool for measuring psychological, emotional and personality features during video interviews (Heilweil 2019). Online interviews have become the norm under COVID-19 lockdowns, and this trend seems set to continue, so the use of AI technology in these contexts may increase. However, tools that interpret facial features may manifest limitations similar to those of recruitment AI, although their impact is not as widely publicised as that of the Amazon case. This means that sustained ethical alertness is required when it comes to preventing violations of the human right to non-discrimination. Or, as a human rights commentator has noted, the problem of "garbage in, garbage out" (Lentz 2021) has to be solved before HR departments can use AI in an ethical manner to substitute human for machine decision-making.

2.2.2 *Case 2: Discriminatory Use of AI in Law Enforcement and Predictive Policing*

> Glenn Rodríguez had been arrested at the age of 16 for his role in the armed robbery of a car dealership, which left one employee dead. In 2016, 25 years later, he applied to the parole board of the Eastern Correctional Facility in upstate New York for early release. He had a model rehabilitation record at the time (Wexler 2017b). Parole was denied. The justification given by the board was that an AI system called COMPAS had predicted him to be "high risk" and the board "concluded that … release to supervision is not compatible with the welfare of society" (Wexler 2017a). The parole board had no knowledge of how the COMPAS risk score was calculated, as the company that had developed the system considered their algorithm a trade secret (ibid). Through cross-referencing with other inmates' scores, Rodríguez found out that the reason for his high-risk score was a subjective personal view given by prison guards, who may have been influenced by racial prejudices. In the end, he was released early. However, "had he been able to examine and contest the logic of the COMPAS system to prove that its score gave a distorted picture of his life, he might have gone home much earlier" (Wexler 2017b)

Rodríguez's case is an example of the discriminatory use of AI in criminal justice, which also includes prominent AI applications for the purposes of predictive policing. "Predictive policing makes use of information technology, data, and analytical techniques in order to identify likely places and times of future crimes or individuals at high risk of [re-]offending or becoming victims of crime." (Mugari and Obioha 2021: 1). The idea behind predictive policing is that existing law enforcement data can improve the targeting of policing interventions. Police resources are limited and it would be desirable to focus them where they are most likely to make a difference, that is, to disrupt or prevent crime or, once crime has been committed, to protect victims, arrest offenders etc. Predictive policing uses past crime data to detect patterns suitable for extrapolation into the future, thereby, one hopes, helping police to identify locations and times when crime is most likely to occur. This is where resources are then deployed.

These ideas sound plausible and are already implemented in many jurisdictions. The most high-profile cases are from the US, where police have been developing and using predictive policing tools in Chicago, Los Angeles, New Orleans and New York since as far back as 2012 (McCarthy 2019). In the UK, research by an NGO showed that "at least 14 UK police forces have used or intend to use … computer algorithms to predict where crime will be committed and by whom" (Liberty n.d.). It is also known that China, Denmark, Germany, India, the Netherlands, and Japan are testing and possibly deploying predictive policing tools (McCarthy 2019).

While the idea of helping the police do their job better, and possibly at reduced cost, will be welcomed by many, the practice of predictive policing has turned out to be ethically problematic. The use of past crime data means that historical patterns are reproduced, and this may become a self-fulfilling prophecy.

For example, areas that historically have high crime rates tend to be those that have lower levels of wealth and educational attainment among the population, as well as higher percentages of migrants or stateless people. Using predictive policing tools means that people who live in deprived areas are singled out for additional police attention, whether they have anything to do with perpetrating any crimes or not. Using algorithmic systems to support policing work has the potential to exacerbate already entrenched discrimination. It is worth pointing out, however, that given awareness of the issue, it is also conceivable that such systems could explicitly screen police activity for bias and help alleviate the problem. The AI systems used for predictive policing and law enforcement could be used to extract and visualise crime data that would make more obvious whether and how crime statistics are skewed in ways that might be linked to ethnic or racial characteristics. This, in turn, would provide a good starting point for a more detailed analysis of the mechanisms that contribute to such developments.

This problem of possible discrimination in relation to specific geographical areas can also occur in relation to individuals. Automated biometric recognition can be used in police cameras, providing police officers with automated risk scores for people they interact with. This then disadvantages people with prior convictions or a past history of interaction with the police, which again tends to over-represent disadvantaged communities, notably those from ethnic minorities. The same logic applies further down the law enforcement chain, when the analysis of data from offenders is used to predict their personal likelihood of reoffending. When the AI tool which informed the decision to hold Glenn Rodríguez in prison for longer than necessary was later examined, it was found that "a disproportionate number of black defendants were 'false positives': they were classified by COMPAS as high risk but subsequently not charged with another crime." (Courtland 2018).

2.2.3 Case 3: Discrimination on the Basis of Skin Colour

In 2016, a 22-year-old engineering student from New Zealand had his passport photo rejected by the systems of the New Zealand department of internal affairs because his eyes were allegedly closed. The student was of Asian descent and his eyes were open. The automatic photo recognition tool declared the photo invalid and the student could not renew his passport. He later told the press very graciously: "No hard feelings on my part, I've always had very small eyes and facial recognition technology is relatively new and unsophisticated" (Reuters 2016). Similar cases of ethnicity-based errors by passport photo recognition tools have affected dark-skinned women in the UK. "Photos of women with the darkest skin were four times more likely to be graded poor quality, than women with the lightest skin" (Ahmed 2020). For instance, a black student's photo was declared unsuitable as her mouth was allegedly open, which it in fact was not (ibid).

Zou and Schiebinger (2018) have explained how such discriminatory bias can occur. As noted earlier, one of the main reasons for discriminatory AI tools is the training sets used.

> Deep neural networks for image classification … are often trained on ImageNet … More than 45% of ImageNet data, which fuels research in computer vision, comes from the United States, home to only 4% of the world's population.

Hence, some groups are heavily over-represented in training sets while others are under-represented, leading to the perpetuation of ethnicity-based discrimination.

2.3 Ethical Questions Concerning AI-Enabled Discrimination

The reproduction of biases and resulting discrimination are among the most prominent ethical concerns about AI (Veale and Binns 2017; Access Now Policy Team 2018). Bias has been described as the "one of the biggest risks associated with AI" (PwC 2019: 13).

The term "discrimination" has at least two distinct meanings, which differ significantly in terms of an ethical analysis (Cambridge Dictionary n.d.). On one hand "discrimination" means the ability to judge phenomena and distinguish between them in a reasonable manner. In this sense, the term has synonyms like "distinction" and "differentiation". For instance, it is a good evolutionary trait for humans to have the ability to distinguish malaria-carrying mosquitoes from flies. The other more widespread contemporary meaning of the term focuses on the unjust or prejudicial application of distinctions made between people, in particular on the basis of their race, sex, age or disability. The former meaning can be ethically neutral, whereas the latter is generally acknowledged to be a significant ethical problem, hence article 7 of the Universal Declaration of Human Rights (see box above). When we use the term "discrimination" in this discussion, we are talking about the ethically relevant type, which is also often illegal.

However, being able to distinguish between phenomena is one of the strengths of AI. Machine-learning algorithms are specifically trained to distinguish between classes of phenomena, and their success in doing so is the main reason for the current emphasis on AI use in a wide field of applications.

AI systems have become increasingly adept at drawing distinctions, at first between pictures of cats and pictures of dogs, which provided the basis for their use in more socially relevant fields, such as medical pathology, where they can distinguish images of cancer cells from those of healthy tissue, or in the business world, where they can distinguish fraudulent insurance claims from genuine ones. The problem is not identifying differences in the broad sense but discrimination on the basis of those particular characteristics.

Unfair/illegal discrimination is a widespread characteristic of many social interactions independent of AI use. While there is broad agreement that job offers should

not depend on an applicant's gender, and that judicial or law enforcement decisions should not depend on a person's ethnicity, it is also clear that they often do, reflecting ingrained systemic injustices. An AI system that is trained on historical data that includes data from processes that structurally discriminated against people will replicate that discrimination. As our case studies have shown, these underlying patterns in the data are difficult to eradicate. Attempts to address such problems by providing more inclusive data may offer avenues for overcoming them. However, there are many cases where no alternative relevant datasets exist. In such cases, which include law enforcement and criminal justice applications, the attempt to modify the data to reduce or eliminate underlying biases may inadvertently introduce new challenges.

However, there are cases where the problem is not so much that no unbiased datasets exist but that the possibility of introducing biases through a poor choice of training data is not sufficiently taken into account. An example is unfair/illegal discrimination arising from poor systems design through a poor choice of training data. Our third case study points in this direction. When the images of 4% of the world population constitute 45% of the images used in AI system design (Zou and Schiebinger 2018), it is reasonable to foresee unfair/illegal discrimination in the results.

This type of discrimination will typically arise when a machine-learning system is trained on data that does not fully represent the population that the system is meant to be applied to. Skin colour is an obvious example, where models based on data from one ethnic group do not work properly when applied to a different group. Such cases are similar to the earlier ones (Amazon HR and parole) in that there is a pre-existing bias in the original data used to train the model. The difference between the two types of discrimination is the source of the bias in the training data. In the first two cases the biases were introduced by the systems involved in creating the data, i.e. in recruitment processes and law enforcement, where women and racial minorities were disadvantaged by past recruiters and past parole boards that had applied structurally sexist or racist perspectives. In the case of discrimination based on skin colour, the bias was introduced by a failure to select comprehensive datasets that included representation from all user communities. This difference is subtle and not always clear-cut. It may be important, however, in that ways of identifying and rectifying particular problems may differ significantly.

Discrimination against people on the basis of gender, race, age etc. is not only an ethical issue; in many jurisdictions such discrimination is also illegal. In the UK, for example, the Equality Act (2010) defines nine protected characteristics: age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, and sexual orientation. Discrimination in the workplace and in wider society based on these protected characteristics is prohibited.

The legal codification of the prohibition of such discrimination points to a strong societal consensus that such discrimination is to be avoided. It raises difficult questions, however, with regard to unfair discrimination that is based on characteristics other than the legally protected ones. It is conceivable that a system would identify

patterns on the basis of other variables that we may not yet even be aware of. Individuals could then be categorised in ways that are detrimental to them. This might not involve protected characteristics, but could still be perceived as unfair discrimination.

Another example of a problematic variable is social class. It is well established that class is an important variable that determines not just individual life chances, but also the collective treatment of groups. Marx's (2017) dictum that the history of all existing society is the history of class struggles exemplifies this position. Discrimination can happen because of a particular characteristic, such as gender, race or disability, but it often happens where individuals combine several of these characteristics that individually can lead to discrimination and, when taken together, exacerbate the discriminatory effect. The term "intersectionality" is sometimes used to indicate this phenomenon (Collins and Bilge 2020). Intersectionality has been recognised as a concern that needs to be considered in various aspects of information technology (IT), not only AI (Fothergill et al. 2019; Zheng and Walsham 2021). It points to the fact that the exact causes of discrimination will in practice often be difficult to identify, which raises questions about the mechanisms of unfair/illegal discrimination as well as ways of addressing them. If the person who is discriminated against is a black, disabled, working-class woman, then it may be impossible to determine which characteristic led to the discrimination, and whether the discrimination was based on protected characteristics and thus illegal.

Hence, unfair/illegal discrimination is not a simple matter. Discrimination based on protected characteristics is deemed to be ethically unacceptable in most democratic states and therefore also typically illegal. But this does not mean that there is no discrimination in social reality, nor should we take it as given that the nature of these protected characteristics will remain constant or that discrimination based on gender, age, race etc. are the only forms of unfair discrimination.

2.4 Responses to Unfair/Illegal Discrimination

With unfair/illegal discrimination recognised as a key ethical problem related to AI and machine learning, there is no shortage of attempts to address and mitigate it. These range from the technical level, where attempts are made to better understand whether training data contains biases that lead to discrimination, to legislative processes where existing anti-discrimination policies are refocused on novel technologies.

One prominent field of research with significant implications regarding unfair/illegal discrimination is that of explainable AI (Holzinger et al. 2017; Gunning et al. 2019). There are many approaches to explainable AI, but what they have in common is an attempt to render the opaque nature of the transformation from input variables to output variables easier to understand. The logic is that an ability to understand *how* an AI system came to a classification of a particular observation would allow the determination of whether that classification is discriminatory and, as a result, could be challenged. If AI is fully explainable, then it should be easy to

see whether gender (sexism) determines employment offers, or whether racism has consequences for law enforcement practices.

While this approach is plausible, it runs into technical and social limits. The technical limits include the fact that machine learning models include large numbers of variables and by their very nature are not easy to understand. If it were possible to reduce them to simple tests of specific variables, then machine learning would not be needed in the first place. However, it might be possible for explainable AI to find ways of testing whether an AI system makes use of protected characteristics and to correct for this (Mittelstadt 2019). Hence, rather than humans making these assessments, another or the same AI system would do so.

When thinking about ways of addressing the role that AI plays in unfair/illegal discrimination, it helps to keep in mind that such discrimination is pervasive in many social processes. Real-life data used for training purposes will often include cases of unfair discrimination and thus lead to their reproduction. Removing traces of structural discrimination from training data, for example by removing data referring to protected characteristics, may not work or may reduce the value of the data for training purposes. The importance of data quality to the trustworthiness of the outcomes of an AI system is widely recognised. The European Commission's proposal for regulating AI, for example, stipulates that "training, validation and testing data sets shall be relevant, representative, free of errors and complete" (European Commission 2021: art. 10(3)). It is not clear, however whether such data quality requirements can possibly be met with real-life data.

Two suggestions on how to address unfair/illegal discrimination (Stahl 2021) will be highlighted here: AI impact assessments and ethics by design.

2.4.1 AI Impact Assessment

The idea of an AI impact assessment is based on the insights derived from many other types of impact assessment, such as social impact assessment (Becker 2001; Becker and Vanclay 2003; Hartley and Wood 2005) and human rights impact assessment (Microsoft and Article One 2018). In general terms, impact assessment aims to come to a better understanding of the possible and likely issues that can arise in a particular field, and use this understanding to prepare mitigation measures. There are several examples of impact assessment that focus on information technologies and topics of relevance to AI, such as privacy/data protection impact assessment (CNIL 2015; Ivanova 2020), ICT ethics impact assessment (Wright 2011) and ethics impact assessment for research and innovation (CEN-CENELEC 2017). The idea of applying impact assessments specifically to AI and using them to get an early warning of possible ethical issues is therefore plausible. This has led to a number of calls for such specific impact assessments for AI by bodies such as the European Data Protection Supervisor (EDPS 2020), UNESCO (2020), the European Fundamental Rights Agency (FRA 2020) and the UK AI council (2021).

The discussion of what such an AI impact assessment should look like in detail is ongoing, but several proposals are available. Examples include the assessment list for trustworthy AI of the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG 2020), the AI Now Institute's algorithmic impact assessment (Reisman et al. 2018), the IEEE's recommended practice for assessing the impact of autonomous and intelligent systems on human wellbeing (IEEE 2020) and the ECP Platform's AI impact assessment (ECP 2019).

The idea common to these AI impact assessments is that they provide a structure for thinking about aspects that are likely to raise concerns at a later stage. They highlight such issues and often propose processes to be put in place to address them. In a narrow sense they can be seen as an aspect of risk management. More broadly they can be interpreted as a proactive engagement that typically includes stakeholder consultation to ensure that likely and foreseeable problems do not arise. Bias and unfair/illegal discrimination figure strongly among these foreseeable problems.

The impact assessment aims to ascertain that appropriate mechanisms for dealing with potential sources of bias and unfair discrimination are flagged early and considered by those designing AI systems. The AI HLEG (2020) assessment, for example, asks whether strategies for avoiding biases are in place, how the diversity and representativeness of end users is considered, whether AI designers and developers have benefitted from education and awareness initiatives to sensitise them to the problem, how such issues can be reported and whether a consistent use of the terminology pertaining to fairness is ensured.

An AI impact assessment is therefore likely to be a good way of raising awareness of the possibility and likelihood that an AI system may raise concerns about unfair/illegal discrimination, and of which form this discrimination might take. However, it typically does not go far in providing a pathway towards addressing such discrimination, which is the ambition of ethics by design.

2.4.2 Ethics by Design

Ethics by design for AI has been developed in line with previous discussions of value-sensitive design (Friedman et al. 2008; van den Hoven 2013). The underlying idea is that an explicit consideration of shared values during the design and development process of a project or technology will be conducive to the embedding of such a value in the technology and its eventual use. The concept has been prominently adopted for particular values, for example in the area of privacy by design (ICO 2008) or security by design (Cavoukian 2017).

A key premise of value-sensitive design is that technology is not a value-neutral tool that can be used for any purposes; design decisions influence the way in which a technology can be used and what consequences such use will have. This idea may be most easily exemplified using the value of security by design. Cybersecurity is generally recognised as an important concern that requires continuous vigilance from individuals, organisations and society. It is also well recognised that some systems are

easier to protect from malicious interventions than others. One distinguishing factor between more secure and less secure systems is that secure systems tend to be built with security considerations integrated into the earliest stages of systems design. Highlighting the importance of security, for example in the systems requirement specifications, makes it more likely that the subsequent steps of systems development will be sensitive to the relevance of security and ensure that the system overall contains features that support security. Value-sensitive design is predicated on the assumption that a similar logic can be followed for all sorts of values.

The concept of ethics by design was developed by Philip Brey and his collaborators (Brey and Dainow 2020) with a particular view to embedding *ethical* values in the design and development of AI and related technologies. This approach starts by highlighting the values that are likely to be affected by a particular technology. Brey and Dainow (2020) take their point of departure from the AI HLEG (2019) and identify the following values as relevant: human agency, privacy and data governance, fairness, wellbeing, accountability and oversight, and transparency. The value of fairness is key to addressing questions of bias and unfair/illegal discrimination.

Where ethics by design goes beyond an ex-ante impact assessment is where it specifically proposes ways of integrating attention to the relevant values into the design process. For this purpose, Brey and Dainow (2020) look at the way in which software is designed. Starting with a high-level overview, they distinguish different design phases and translate the ethical values into specific objectives and requirements that can then be fed into the development process. They also propose ways in which this can be achieved in the context of agile development methodologies. This explicit link between ethical concerns and systems development methodologies is a key conceptual innovation of ethics by design. Systems development methodologies are among the foundations of computer science. They aim to ensure that systems can be built according to specifications and perform as expected. The history of computer science has seen the emergence of numerous design methodologies. What Brey and his colleagues have done is to identify universal components that most systems development methodologies share (e.g. objectives specification, requirements elicitation, coding, testing) and to provide guidance on how ethical values can be integrated and reflected in these steps.

This method has only recently been proposed and has not yet been evaluated. It nevertheless seems to offer an avenue for the practical implementation of ethical values, including the avoidance of unfair/illegal discrimination in AI systems. In light of the pervasive nature of unfair/illegal discrimination in most areas of society one can safely say that all AI systems need to be built and used in ways that recognise the possibility of discrimination. Failure to take this possibility into account means that the status quo will be reproduced using AI, which will often be neither ethical nor legal.

2.5 Key Insights

Unfair/illegal discrimination is not a new problem, nor one that is confined to technology. However, AI systems have the proven potential to exacerbate and perpetuate it. A key problem in addressing and possibly overcoming unfair/illegal discrimination is that it is pervasive and often hidden from sight. High-profile examples of such discrimination on the basis of gender and race have highlighted the problem, as in our case studies. But unfair/illegal discrimination cannot be addressed by looking at technology alone. The broader societal questions of discrimination need to be considered.

One should also not underestimate the potential for AI to be used as a tool to *identify* cases of unfair/illegal discrimination. The ability of AI to recognise patterns and process large amounts of data means that AI may also be used to demonstrate where discrimination is occurring.

It is too early to evaluate whether – and, if so, how far – AI impact assessment will eliminate the possibility of unfair/illegal discrimination through AI systems. In any event, discrimination on the basis of protected characteristics requires access to personal data, which is the topic of the next chapter, on privacy and data protection.

References

Access Now (2018) Human rights in the age of artificial intelligence. Access Now. https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf. Accessed 1 May 2022

Access Now Policy Team (2018) The Toronto declaration: protecting the right to equality and non-discrimination in machine learning systems. Access Now, Toronto. https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf. Accessed 26 Sept 2020

Ahmed M (2020) UK passport photo checker shows bias against dark-skinned women. BBC News, 8 Oct. https://www.bbc.com/news/technology-54349538. Accessed 4 May 2022

AI HLEG (2019) Ethics guidelines for trustworthy AI. High-level expert group on artificial intelligence. European Commission, Brussels. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419. Accessed 25 Sept 2020

AI HLEG (2020) The assessment list for trustworthy AI (ALTAI). High-level expert group on artificial intelligence. European Commission, Brussels. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342. Accessed 10 Oct 2020

Becker HA (2001) Social impact assessment. Eur J Oper Res 128:311–321. https://doi.org/10.1016/S0377-2217(00)00074-6

Becker HA, Vanclay F (eds) (2003) The international handbook of social impact assessment: conceptual and methodological advances. Edward Elgar Publishing, Cheltenham

Brey P, Dainow B (2020) Ethics by design and ethics of use approaches for artificial intelligence, robotics and big data. SIENNA. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

Cambridge Dictionary (n.d.) Discrimination. https://dictionary.cambridge.org/dictionary/english/discrimination. Accessed 4 May 2022

Cavoukian A (2017) Global privacy and security, by design: turning the 'privacy vs. security' paradigm on its head. Health Technol 7:329–333. https://doi.org/10.1007/s12553-017-0207-1

CEN-CENELEC (2017) Ethics assessment for research and innovation, part 2: ethical impact assessment framework. CWA 17145-2. European Committee for Standardization, Brussels. http://ftp.cencenelec.eu/EN/ResearchInnovation/CWA/CWA17214502.pdf. Accessed 6 Oct 2020

CNIL (2015) Privacy impact assessment (PIA): methodology. Commission Nationale de l'Informatique et des Libertés, Paris

Collins PH, Bilge S (2020) Intersectionality. Wiley, New York

Courtland R (2018) Bias detectives: the researchers striving to make algorithms fair. Nature 558:357–360. https://doi.org/10.1038/d41586-018-05469-3

ECP (2019) Artificial intelligence impact assessment. ECP Platform for the Information Society, The Hague. https://ecp.nl/wp-content/uploads/2019/01/Artificial-Intelligence-Impact-Assessment-English.pdf. Accessed 1 May 2022

EDPS (2020) EDPS opinion on the European Commission's white paper on artificial intelligence: a European approach to excellence and trust (opinion 4/2020). European Data Protection Supervisor, Brussels. https://edps.europa.eu/data-protection/our-work/publications/opinions/edps-opinion-european-commissions-white-paper_en. Accessed 6 May 2022

Equality Act (2010) c15. HMSO, London. https://www.legislation.gov.uk/ukpga/2010/15/contents. Accessed 5 May 2022

European Commission (2021) Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. European Commission, Brussels. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. Accessed 1 May 2022

Fothergill BT, Knight W, Stahl BC, Ulnicane I (2019) Intersectional observations of the Human Brain Project's approach to sex and gender. J Inf Commun Ethics Soc 17:128–144. https://doi.org/10.1108/JICES-11-2018-0091

FRA (2020) Getting the future right: artificial intelligence and fundamental rights. European Union Agency for Fundamental Rights, Luxembourg

Friedman B, Kahn P, Borning A (2008) Value sensitive design and information systems. In: Himma K, Tavani H (eds) The handbook of information and computer ethics. Wiley Blackwell, Hoboken, pp 69–102

Gunning D, Stefik M, Choi J et al (2019) XAI: explainable artificial intelligence. Sci Robot 4(37). https://doi.org/10.1126/scirobotics.aay7120

Hartley N, Wood C (2005) Public participation in environmental impact assessment: implementing the Aarhus convention. Environ Impact Assess Rev 25:319–340. https://doi.org/10.1016/j.eiar.2004.12.002

Heilweil R (2019) Artificial intelligence will help determine if you get your next job. Vox-Recode, 12 Dec. https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen. Accessed 4 May

Holzinger A, Biemann C, Pattichis CS, Kell DB (2017) What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923 [cs, stat]. https://doi.org/10.48550/arXiv.1712.09923

ICO (2008) Privacy by design. Information Commissioner's Office, Wilmslow. https://web.archive.org/web/20121222044417if_/http://www.ico.gov.uk:80/upload/documents/pdb_report_html/privacy_by_design_report_v2.pdf. Accessed 6 Oct 2020

IEEE (2020) 7010-2020: IEEE recommended practice for assessing the impact of autonomous and intelligent systems on human well-being. IEEE Standards Association, Piscataway, NJ. https://doi.org/10.1109/IEEESTD.2020.9084219

Ivanova Y (2020) The data protection impact assessment as a tool to enforce non-discriminatory AI. In: Antunes L, Naldi M, Italiano GF et al (eds) Privacy technologies and policy. 8th Annual privacy forum, APF 2020, Lisbon, Portugal, 22–23 Oct. Springer Nature Switzerland, Cham, pp 3–24. https://doi.org/10.1007/978-3-030-55196-4_1

Kaur D (2021) Has artificial intelligence revolutionized recruitment? Tech Wire Asia, 9 Feb. https://techwireasia.com/2021/02/has-artificial-intelligence-revolutionized-recruitment/. Accessed 4 May 2022

Latonero M (2018) Governing artificial intelligence: upholding human rights & dignity. Data & Society. https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf. Accessed 26 Sept 2020

Lentz A (2021) Garbage in, garbage out: is AI discriminatory or simply a mirror of IRL inequalities? 18 Jan. Universal Rights Group, Geneva. https://www.universal-rights.org/blog/garbage-in-garbage-out-is-ai-discriminatory-or-simply-a-mirror-of-irl-inequalities/. Accessed 4 May 2022

Liberty (n.d.) Predictive policing. https://www.libertyhumanrights.org.uk/fundamental/predictive-policing/. Accessed 4 May 2022

Marx K (2017) Manifest der Kommunistischen Partei. e-artnow

McCarthy OJ (2019) AI & global governance: turning the tide on crime with predictive policing. Centre for Policy Research, United Nations University. https://cpr.unu.edu/publications/articles/ai-global-governance-turning-the-tide-on-crime-with-predictive-policing.html. Accessed 4 May 2022

Microsoft, Article One (2018) Human rights impact assessment (HRIA) of the human rights risks and opportunities related to artificial intelligence (AI). https://www.articleoneadvisors.com/case-studies-microsoft. Accessed 1 May 2022

Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1:501–507. https://doi.org/10.1038/s42256-019-0114-4

Mugari I, Obioha EE (2021) Predictive policing and crime control in the United States of America and Europe: trends in a decade of research and the future of predictive policing. Soc Sci 10:234. https://doi.org/10.3390/socsci10060234

Muller C (2020) The impact of artificial intelligence on human rights, democracy and the rule of law. Ad Hoc Committee on Artificial Intelligence (CAHAI), Council of Europe, Strasbourg. https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da. Accessed 2 May 2022

PwC (2019) A practical guide to responsible artificial intelligence. https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide.pdf. Accessed 18 June 2020

Reisman D, Schultz J, Crawford K, Whittaker M (2018) Algorithmic impact assessments: a practical framework for public agency accountability. AI Now Institute, New York. https://ainowinstitute.org/aiareport2018.pdf. Accessed 18 June 2020

Reuters (2016) Passport robot tells Asian man his eyes are closed. New York Post, 7 Dec. https://nypost.com/2016/12/07/passport-robot-tells-asian-man-his-eyes-are-closed/. Accessed 4 May 2022

Reuters (2018) Amazon ditched AI recruiting tool that favored men for technical job. The Guardian, 11 Oct. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine. Accessed 4 May 2022

Stahl BC (2021) Artificial intelligence for a better future: an ecosystem perspective on the ethics of AI and emerging digital technologies. Springer Nature Switzerland AG, Cham. https://doi.org/10.1007/978-3-030-69978-9

UK AI Council (2021) AI roadmap. Office for Artificial Intelligence, London. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/949539/AI_Council_AI_Roadmap.pdf

UN (1948) Universal declaration of human rights. http://www.un.org/en/universal-declaration-human-rights/. Accessed 4 May 2022

UNESCO (2020) First draft text of the recommendation on the ethics of artificial intelligence, 7 Sept. Ad hoc expert group (AHEG) for the preparation of a draft text, UNESCO, Paris. https://unesdoc.unesco.org/ark:/48223/pf0000373434. Accessed 12 Oct 2020

van den Hoven J (2013) Value sensitive design and responsible innovation. In: Owen R, Heintz M, Bessant J (eds) Responsible innovation. Wiley, Chichester, pp 75–84

Veale M, Binns R (2017) Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. Big Data Soc 4(2). https://doi.org/10.1177/2053951717743530

Wexler R (2017a) Code of silence. Washington Monthly, 11 June. https://washingtonmonthly.com/2017a/06/11/code-of-silence/. Accessed 4 May 2022

Wexler R (2017b) When a computer program keeps you in jail. The New York Times, 13 June. https://www.nytimes.com/2017b/06/13/opinion/how-computers-are-harming-criminal-justice.html. Accessed 4 May 2022

Wright D (2011) A framework for the ethical impact assessment of information technology. Ethics Inf Technol 13:199–226. https://doi.org/10.1007/s10676-010-9242-6

Zheng Y, Walsham G (2021) Inequality of what? An intersectional approach to digital inequality under Covid-19. Inf Organ 31:100341. https://doi.org/10.1016/j.infoandorg.2021.100341

Zou J, Schiebinger L (2018) AI can be sexist and racist: it's time to make it fair. Nature 559:324–326. https://doi.org/10.1038/d41586-018-05707-8