6-2024

AI employment decision-making: Integrating the equal opportunity merit principle and explainable AI

Gary Kok Yew CHAN
Singapore Management University, garychan@smu.edu.sg

AI & SOCIETY (2024) 39:1027–1038
https://doi.org/10.1007/s00146-022-01532-w

OPEN FORUM

AI employment decision-making: integrating the equal opportunity merit principle and explainable AI

Gary K Y Chan[1]

Received: 29 September 2021 / Accepted: 20 June 2022 / Published online: 10 July 2022

Abstract

Artificial intelligence (AI) tools used in employment decision-making cut across the multiple stages of job advertisements, shortlisting, interviews and hiring, and actual and potential bias can arise in each of these stages. One major challenge is to mitigate AI bias and promote fairness in opaque AI systems. This paper argues that the equal opportunity merit principle is an ethical approach for fair AI employment decision-making. Further, explainable AI can mitigate the opacity problem by placing greater emphasis on enhancing the understanding of reasonable users (employing organisations) and affected persons (employees and job candidates) as to the AI output. Both the equal opportunity merit principle and explainable AI should be integrated in the design and implementation of AI employment decision-making systems so as to ensure, as far as possible, that the AI output is arrived at through a fair process.

Keywords Artificial intelligence · Employment decision-making · Bias · Fairness · Equal opportunity · Merit · Explainable AI

1 Introduction

In the race for global talent by corporations and organisations, AI can provide specific advantages in employment decision-making. Organisations may utilise AI to reduce costs and time expended on human resources acquisition and to enhance management processes. Machine learning can improve the predictions of worker productivity e.g., in police hiring decisions and teacher tenure decisions (Chalfin et al. 2016). AI offers the prospect of greater consistency in employment decision-making without having to be concerned about human fatigue and aberrations.

The application of AI systems cut across a wide range of employment stages including job advertisements, matching tools, screening of candidates, hiring decisions and even serving as career coaches. Through social media platforms, AI can deliver job advertisements to targeted audiences and enable employing organisations to personalise recruitment and match opportunities to the right candidates. They screen resumes to extract relevant job skills and match applications to jobs with the right fit and scour information from the data mining of candidate profiles using semantic analysis and natural language processing. AI can predict the extent of close matching between the applicant's resume and employers' requirements, and compare the candidates with existing successful employees.

Other functions include administering pre-employment tests, conducting interviews and grading the candidates' responses against interview answers from current successful employees and even analysing facial expressions, eye contact, voice and word choices[1] using emotion recognition software. Data points indicating absenteeism, salary increase, rate of promotion and birth of a child have also been utilised to predict the resignation risks of employees.

Notwithstanding the wide-ranging applications, the use of AI in employment decision-making has attracted controversy due to allegations of bias. Amazon's computer models were trained to vet applicants by observing patterns in

✉ Gary K Y Chan
garychan@smu.edu.sg

[1] Yong Pung How School of Law, Singapore Management University, Singapore, Singapore

[1] Miranda Bogen and Aaron Rieke, "Help Wanted: An Examination of Hiring Algorithms, Equity and Bias", December 2018 at p. 35; see also https://www.inc.com/minda-zetlin/ai-is-now-analyzing-candidates-facial-expressions-during-video-job-interviews.html.

resumes submitted to the company over a 10-year period. However, in 2018, Amazon decided to abandon the use of AI for screening job applicants.[2] The resumes were mainly from men who dominated the tech industry, and the AI gave less emphasis to those that included the word "women's" and downgraded the graduates of two all-women's colleges.

In job advertisements, search engines may deliver job postings on well-paying technical jobs that are targeted at men only, possibly discriminating against women job-seekers.[3] Bias may also originate from the lexical and semantic differences that exist in the text of resumes distinguishing different genders. As AI system track users' interests based on their clicks and actions (content filtering) and what other people similar to the users are interested in (collaborative filtering), it can reinforce the users' cognitive biases (Bogen and Rieke 2018, p. 21). Further, questions have been raised about AI-driven facial recognition technology and biometric data in pre-employment screening and hiring decisions.[4]

Unless properly justified in the eyes of the users and members of the public, AI bias can adversely affect public and social trust on a wide scale. Employment decisions are ubiquitous in impact. Most of us have at some point in our lives been subject to or affected by employment decisions or have been in a position to make employment decisions on behalf of organisations. Where there is grave bias arising from the use of AI and/or a knowledge gap as to fairness in employment decision-making between AI designers and the human resource practitioners, employees and job candidates, trust in (or reliance on) AI systems (Ryan 2020) could be seriously impacted.

Bias mitigation measures, whilst important to maintain such trust or reliance, presuppose a rubric for fairness applicable to AI systems. However, the search for a generally accepted fairness rubric has been elusive. Outside of the AI domain, there are several plausible conceptions of substantive fairness. In view of the heterogeneity in the concept of fairness, it is challenging for developers and users of AI to determine and implement a uniform and consistent fairness requirement for AI.

Additionally, AI systems such as artificial neural networks may render the functioning of the systems and their outputs unduly complex to users and, at times, even to AI experts. When AI systems are opaque, they invite doubts as to whether the processes in which the AI decisions have been generated and/or the decisions themselves have been fair to job candidates and employees (Ajunwa 2020a).

This paper argues that an ethical approach on access to job opportunities embodied in what I would refer to as the equal opportunity merit principle offers great promise for application to the specific employment context. To counter the problem of opacity of AI systems, there should be more emphasis on the nature and extent of understanding of AI users and affected persons applicable to the employment decision-making context beyond the explanation of AI models, processes and design. Furthermore, the abovementioned explainable AI approach would be capable of supporting the equal opportunity merit principle with a view to promoting fairness in AI employment decision-making.

To begin the examination, the problem of AI bias will first be described in the next section with the aid of a diagrammatic model specially catered to the employment context. We will then focus on the meaning and applicability of the equal opportunity merit principle to employment decision-making amidst the controversies surrounding the concept of fairness. The next section considers the challenges posed by the opacity of AI systems followed by the proposed person-centric perspective to explainable AI in connection with the different types of explanation. Finally, we will discuss how the equal opportunity merit principle can be integrated with the concept and practice of explainable AI in the design and implementation of AI employment decision-making.

2 The problem of AI bias in employment

As a starting point, bias can potentially arise when persons in similar circumstances are treated differently or conversely, when persons in different circumstances are treated the same. The mere difference of treatment of particular groups does not however constitute bias. This is because the difference in treatment could be due to and justified by the different attributes (e.g., skills and knowledge) belonging to the group members and/or a difference in circumstances to which the group members are subject (e.g., poverty). Bias arises only where there is disproportionate treatment or outcomes that go beyond justifiable differences in attributes or circumstances.

The assessment of bias is twofold based on the differential treatment intentionally meted out to certain groups and/or the differential outcomes or impact generated. Insofar as AI bias is concerned, it is in essence a "systematic error" privileging certain groups at the expense of others resulting in disadvantages or harms to the latter (Altman et al. 2018; Bellamy et al. 2018). In the employment context, the disadvantages and harms may be assessed on both the individual and

[2] Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women" Business News, 10 October 2018 at https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[3] "ACLU Says Facebook Ads let Employers Favor Men Over Women", WIRED, 18 Sept 2018.

[4] https://www.forbes.com/sites/patriciagbarnes/2020/02/03/group-asks-federal-trade-commission-to-regulate-use-of-artificial-intelligence-in-pre-employment-screenings/#7930fa932b54, and https://epic.org/privacy/ftc/hirevue/.

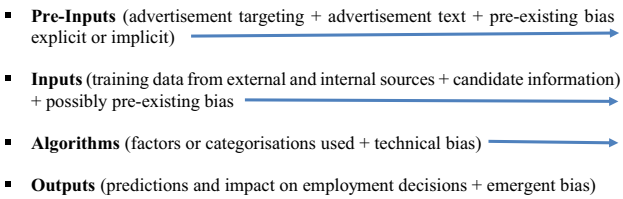
societal levels. First, the candidate is deprived of the opportunity to enhance his or her potential capacities in the job. More tangibly, bias can result in direct adverse impact on the candidate's income and livelihood if he or she is unfairly excluded from employment and indirectly, on the candidate's dependants. The outcome need not necessarily be the loss of a specific job that the candidate would otherwise be entitled to but can include the loss of job prospects in the market by virtue of being unfairly excluded. From the wider societal perspective, AI bias can amplify and entrench community prejudices and stereotypes and undermine a system based on meritocracy.

Friedman and Nissenbaum (1996) referred to three types of bias in computer systems: (i) pre-existing bias (i.e., bias having roots in social institutions, practices and attitudes and the personal bias of clients or system designers whether explicit or implicit); (ii) technical bias (i.e., software, hardware and technical limitations of computer systems or algorithms); and (iii) emergent bias (which arises in the context of the use of computer systems with real users typically after the design is completed as a result of changing societal knowledge, population and cultural values).

For purposes of this paper, we will adopt the abovementioned classification of pre-existing, technical and emergent bias as applied to the employment context. Three additional points may be mentioned here. First, emergent bias can arise from the use of computer systems based on the algorithms designed and not necessarily from changes in societal knowledge and values. For example, algorithmically-designed chatbots may learn from prevailing human prejudices and stereotypical views and adopt racist language against particular groups of people. In the employment context, the chatbots for conducting interviews may learn from existing human prejudices through conversations. Second, another aspect of emergent bias known as "automation bias" may arise when humans give undue weight to predictions and scores by automated means. Third, technical bias need not be confined to computer or algorithmic limitations. Human choice (and bias) in the design of algorithms such as in the categorisation process (e.g., whether an algorithmic criterion is job-related or not) can also occur.

The scope of bias is also influenced by legal and ethical considerations. The IEEE P7003 on Algorithmic Bias Considerations[5] defined "negative bias" as the "usage of overly subjective or uniformed data sets or information known to be inconsistent with legislation concerning certain protected characteristics (such as race, gender, sexuality, etc.)", and alternatively, "diminishing stakeholder or user wellbeing" which are regarded as "inappropriate." Hence, the scope of bias can overlap with both the legal stipulations and ethical standards based on stakeholder and societal well-being that may vary from country to country.

Bearing in mind the extent of AI use in employment decision-making and the discussion of bias above, the following diagrammatic model for assessing the sources of AI bias is proposed:



Bias (human and algorithmic) can occur at each stage from Pre-inputs to Outputs (cf the approach in Bellamy et al. 2018 for making fair predictions according to the stages of "fair pre-processing, fair in-processing, and fair post-processing" though it was not specifically catered to the employment context).

At the pre-Input stage, bias may originate from external sources such as the pre-existing discrimination or prejudices in society or internally within the employing organisations and/or in the design of job advertisements which may then skew the applications that are received by the employing organisation. Apart from explicit or conscious bias, there is also implicit human bias associated with discriminatory behaviour (Greenwald and Banaji 1995; Greenwald and Krieger 2006).

At the Input stage, the potential inaccuracies or non-representativeness of the training data from the company itself or from external sources such as Linkedin and social media platforms constitute one potential source of bias. There may also be mislabeling of the data based on the employer's prejudiced interests in favour of certain groups of candidates which may in turn influence the recommendations offered by the AI system (Barocas and Selbst 2016, pp. 682–683). Moreover, the selection of people in the training data may be biased (e.g., the data does not include certain marginalised groups) or the selection of the attributes of the people are incomplete (e.g., where it is difficult to collect data such as personalities suitable for specific jobs) (Calders and Zliobaite 2013). Another source of bias would be the collation of data from certain online sites or social media platforms that may skew the data in favour of those individuals using such sites and platforms (Barocas and Selbst 2016, p. 686).

In the design of algorithms for employment decision-making, the factors and classifications used by the AI designers in consultation with the human resource practitioners are significant. Bias arises when the hiring of employees is not

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

based on objective merit-based factors (which will be discussed in greater detail below).

The multi-staged model mentioned above reminds us that first, the concept of AI bias goes beyond mere algorithmic bias; and secondly, bias at any of the stages will contribute to the overall bias and may even be amplified by the large-scale use of AI. Thus, we should not only be concerned with algorithmic bias per se but also the potential impact of AI in general in maintaining and/or contributing to existing biases.

3 AI fairness and the equal opportunity merit principle

In connection with the pre-existing bias inherent in society mentioned in the previous section, we observe that the assessment of whether existing distributions (of resources, opportunities or capabilities) are fair is not strictly atemporal. Such an assessment would inevitably have to take into account the historical and socio-economic background and circumstances that led to the existing distributions, for example, the underlying bases for the protections for certain groups due to discrimination in the past, the "life experiences" of individuals and group members and their relationships inter se (Binns 2018). However, the complex policy questions as to whether a society should implement affirmative actions to correct or ameliorate past gender and racial discrimination (Davis et al. 2021) and, if so, the nature and extent of such policy implementation, the people's reactions to and perceptions of fairness in policy implementation (Harrison et al. 2006; Sinclair & Carlsson 2021) are beyond the scope of this paper.

Nonetheless, in the arena of employment decision-making by employing organisations in practice, AI can play an important role. First, it can avoid "unconscious" human bias and "noise" from the inconsistencies and subjectivity in human decision-making in hiring (Houser 2019) and enhance the accuracy and validity of data concerning the candidates (Chamorro-Premuzic & Akhtar 2019). As we will see below, AI systems can exclude irrelevant factors such as gender and race in hiring decisions. In addition, technology companies have developed a number of AI bias mitigation and algorithmic fairness tools (Raghavan et al., 2019). They include tools that design the text of job descriptions to ensure the diversity and gender balance of the applicants and remove irrelevant data that may lead to bias in order to shortlist suitable candidates regardless of gender or ethnicity. Open source software toolkits incorporate fairness metrics for datasets and models, bias detection and bias mitigation algorithms (Bellamy et al. 2018). Biases are monitored at multiple stages whether in the initial training data, in the algorithm, or in the predictions made by the classifier, and bias mitigation algorithms are used to improve the fairness metrics.

Furthermore, we can work towards improving the fairness of AI process with an appropriate ethical design and implementation. In this section, we will focus on fairness as amongst employees or job candidates inter se and not fairness between employers and employees or candidates. The literature has shown that there is little consensus as to what amounts to substantive fairness and this can affect the question of the appropriate measures for implementing algorithmic fairness.

One notion that is pivotal to fairness is that of equality, in particular, the question "equality of what?". Should we focus on equality of resources, opportunities or capabilities? These equality parameters, so to speak, can result in different consequences and trade-offs for employing organisations, employees or job candidates. In particular, the implementation of fairness algorithms presents a formidable challenge due to competing approaches to the concept of fairness (Lee et al. 2021).

Dworkin (2000) supports equality of resources based on two fundamental principles, first, that all human lives should flourish, and second, that each person is responsible for defining and achieving flourishing in his or her own life. Hence, true equality means equality in the value of the resources that each person commands, not his or her achievements.

The doctrine of equality of resources attempts to achieve a measure of equality which it cannot never fulfil. This is because resources matter to an individual only to the extent that he values those resources. The values assigned to a similar set of resources can vary significantly amongst individuals according to their individual life-plans and prevailing circumstances. Moreover, ensuring equality of resources is not typically the responsibility of employing organisations.

On the other hand, equality of opportunities demands, first and foremost, equal and open access. There should not be barriers to entry or restrictions against gaining access to opportunities for some groups and not others unless justified. Rawls' notion of "fair equality of opportunity" implies that everyone should have a fair chance to attain public offices and social positions such that those "who have the same level of talent and ability and the same willingness to use these gifts should have the same prospects of success" (Rawls 2001, pp. 43–4). This principle, as part of a larger set of Rawlsian principles,[6] is derived from a hypothetical experiment where social actors with primary goods[7] and a

[6] The other principles are an individual's claim to a set of equal basic liberties and the difference principle that socioeconomic inequalities are for the greatest benefit of the least advantaged members of society: Rawls (2001, p. 42).

[7] The Rawlsian set of primary goods includes rights, liberties, income, opportunities and wealth.

rational plan of life decide on societal rules under a veil of ignorance in which they are not cognisant of their particular talents, abilities or socioeconomic circumstances.

The "powers and prerogatives of offices and positions of authority" under the fair equality of opportunity are primary goods in his schema (Rawls 2001, 58). The significance of opportunities for meaningful work is undergirded by self-respect which is also a primary good (Rawls 1971, 92, 290 & 440; Rawls 1999, 50; Rawls 2001, 59). Thus, the principle of equality of opportunity allows for the self-determination of individuals and respects the personal choice of individuals in pursuing meaningful goals (including jobs) based on their efforts and knowledge or skill acquisition.

One version of the equal opportunity principle is tied to merit (the equal opportunity merit principle). Unequal barriers and access to employment cannot be permitted unless the differential treatment can be justified based on job-relevant attributes (Roemer 2000). Thus, determining job-related merits is the key to understanding and applying the equal opportunity merit principle. As an illustration, a female should have an equal opportunity to obtain a technology job as a male. Equality of opportunity depends ultimately on the ground of distinction; it is perfectly fine to distinguish based on technical skills for the job but not on gender per se (Holmes 2005, p. 192). Thus, the equal opportunity merit principle is an expression of and consistent with the more general principle of equal treatment in similar circumstances mentioned above.

In comparison to equality of resources, Amartya Sen's capabilities approach focuses instead on the freedom to achieve beings and functionings in life (that is, to be somebody or to do something) taking into account our cultural conditions and environment. It is essentially the freedom to achieve what we value (Sen 1992, p. 31). In other words, it is what we are enabled to do with the available resources that is important and not so much the resources themselves (Arneson 1989, pp. 90–1). On the other hand, equality of capabilities are likened to "well-being freedom" which is associated with the notion of equality of opportunities to welfare (Arneson 1989, p. 91). On this same question of "equality of what?", instead of "equal opportunities" to welfare, Cohen (2011, pp. 4 & 14) preferred the term "equal access to advantage" which is wider than welfare and would presumably include the benefits that a job offers. It can arguably be extended to employment benefits such as job transfers, promotions, wage raises or bonuses.

In response to Sen's proposal to focus more on capabilities rather than goods, Rawls argued that primary goods do take into account the basic capabilities of citizens in their exercise of the two moral powers (i.e., the capacity for a sense of justice as well as the capacity for a conception of the good) as fully cooperating members of a society (Rawls 2001, 169). Moreover, these moral powers are supported by the equal liberties enjoyed by the citizens (Rawls 2001, 175). Hence, Rawls' primary goods and a person's basic capabilities are closely inter-linked.

In reality, there is no level playing field in terms of natural endowments as noted by Rawls. In this regard, the equal opportunity merit principle cannot on its own eradicate all societal inequalities that may arise from differences in natural abilities. Under this principle, people may be entitled to their income and position but we cannot say that they deserve them in the moral sense (Sandel 2021, p. 141). Nonetheless, the principle is beneficial in terms of enabling and empowering people to choose their rational plan for life and to be responsible for the outcomes arising from their choices (Temkin 2016, p. 263). Beyond the equal opportunity merit principle, Temkin referred to the alternative basis for equality of the value of life prospects assessed ex ante but there are, admittedly, practical difficulties in implementing this philosophical ideal in practice (Temkin 2016, p. 269).

How can the concept of merit be incorporated in employment decision-making? The task of identifying appropriate job-related factors is by no means simple. To apply the concept of merit, one objective measure may be to select job-related factors that are directly related to the designated job scope. This can include the candidate's technical abilities and qualifications, prior industry experience and skills-set relevant to the job scope or description. Rawls' equal opportunity principle assumes that the criteria for public offices and social positions are known to all eligible candidates. When these candidates apply for the available job vacancies, employing organisations should examine the attributes of candidates in line with the advertised criteria for jobs.

At times, irrelevant factors may be indirectly used as proxies for a job requirement. For example, it would be contrary to merit for an organisation to require candidates to be from a particular race (e.g., Chinese race) when the job scope only called for employees be proficient in a language (e.g., Chinese language) albeit one that is widely spoken in the country where the employees are engaged to serve customers.

A broader criterion is corporate success which may depend on the desired objectives of the employing organisation. A "good employee" may correspond with measurable outcomes such as the company's "relatively higher sales, shorter production time, or longer tenure" (Barocas and Selbst 2016, p. 679). Alternatively, we can refer to classifiers that assess the candidates' "fit" with the corporate culture of the organisation. To ascertain "fit", a more detailed profile of the candidates going beyond qualifications, experience and skills-set is required. This may require evidence of the candidates' unique personality, moral values or character for which the assessment would typically be more subjective.

Relatedly, in theory at least, irrelevant factors should ideally be removed or excluded from the classifiers. One

practical problem that can arise here is that the excluded attributes may be implicit in the non-excluded ones (Romei and Ruggieri 2014, p. 39). This can result in disparate impact on certain disadvantaged groups e.g., certain factors such as job tenure may add to the existing discrimination against female job applicants (Kroll et al. 2017, p. 681). Special attention must be paid by developers to guard against such implicit algorithmic bias when designing algorithms.

The nature and type of fairness, in particular group or individual fairness, may be exemplified in the different approaches to designing algorithms. Corbett-Davies and Goel (2018) referred to the anti-classification approach where protected attributes (e.g., race and gender and their proxies) are not explicitly used to make decisions. This approach appears consistent with the Rawlsian notion of veil of ignorance to omit certain attributes or variables so as to attain procedural justice as fairness. AI that is used to shortlist candidates may conceal certain candidate attributes. To ensure fairness, factors outside an individual's control such as their perceived race or where they were born should be removed as far as possible in line with the equal opportunity merit principle. Yet we should exercise caution where excluding protected attributes such as gender from the AI model can affect the predictive analysis and end up discriminating a particular gender (Corbett-Davies and Goel 2018).

Dwork et al. (2012) argued that statistical parity[8]—a feature of group fairness—can produce unfair outcomes from the perspective of an individual. They advocated instead individual fairness i.e., "the principle that any two individuals who are similar with respect to a particular task should be classified similarly". This is consistent with the principle of equal treatment of persons in similar situations. In similar vein, Corbett-Davies and Goel (2018) contended that it is often preferable to treat similarly risky people (e.g., risks of loan defaults) similarly based on the available statistically accurate estimates of risk.

Subjective characteristics that are not susceptible to easy categorisation may be taken into account as long as they can be applied consistently and non-arbitrarily. Tambe et al. (2019, p. 32) noted that individuals should be acknowledged for their performance-enhancing characteristics (e.g., grit or intrinsic motivation) independent of group membership. To the extent that these qualities of a candidate are demonstrated to be objectively connected to the job performance, they can be taken into account in employment decision-making provided they are also capable of being incorporated as the AI inputs as part of the design and implementation process (see Sect. 5 below).

In sum, the equal opportunity merit principle premised on equal access to job opportunities open to all eligible candidates based on Rawls' theory offers an ethical approach to employment decision-making that takes into consideration socioeconomic realities. Coupled with the merit-based factors, the principle is also capable of being broken down into more concrete components for analysis and application to specific employment contexts. The principle relies on the concept of equality applied in a consistent and non-arbitrary fashion based on largely objective merit-based job factors even if there might be differences in value judgments for the exceptional cases.

That said, there are two limitations relating to the equality opportunity merit principle. First, the existence of disparities in natural endowments distributed across a society leading to unequal outcomes is not incompatible with the equal opportunity merit principle. Second, the discussion on the equal opportunity merit principle, which seeks to achieve equality and objectivity in employment decision-making, is not meant to directly address or correct historical gender and racial discrimination via affirmative action policies.

4 The opacity problem and explainable AI for employment decision-making

Within the discussion on AI bias mitigation and algorithmic fairness is an implicit assumption that we have adequate knowledge about how employment decisions are arrived at. This assumption may be somewhat misplaced with respect to AI outputs and processes. The problem of AI opacity is real and has to be addressed even though the path ahead is neither obvious nor straightforward. We will examine how the problem of AI opacity can be mitigated based on a concept of explainability that is stakeholder-driven, contextualised for employment decision-making in practice and, importantly, supportive of the equal opportunity merit principle.

Opacity at its core is a problem of "mismatch" between the operations of machine learning algorithms and human interpretations (Burrell 2016, p. 3). Two features of opacity are inscrutability and nonintuitiveness (Selbst and Barocas 2018) which inhibit human understanding. In practice, there is a continuum or scales of interpretability from highly interpretable AI models—linear, monotonic functions[9] to low interpretability using nonlinear and nonmonotoic functions (Hall and Gill 2018). AI opacity, to the extent that the parameters, data or process for reaching a particular AI output are unknown, may compound the problem of

[8] This means that "the demographics of the set of individuals receiving any classification are the same as the demographics of the underlying population".

[9] The terms "linear, monotonic" means "for a change in any given input variable (or sometimes combination or function of an input variable), the output of the response function changes at a defined rate,

discrimination against groups or individuals (Heinrichs 2021). The discrimination and adverse effects arising from such epistemic opacity may occur unintentionally and continue undetected for long periods.

The use of opaque AI to generate output for employment decision-making has been challenged in the courts. In *Houston Federation of Teachers v Houston Independent School District*,[10] algorithms were used to make a decision to terminate teachers in public schools based on student performance in standardised tests. The focus was the 14th Amendment which states that a person should not be deprived of life, liberty and property without due process. The Federation of Public Teachers sought declaratory judgment and injunction to restrain the use of the scores to terminate employment, arguing that they were denied access to the algorithms and data to verify the accuracy of the scores. The court noted that such scores may be inaccurate, and the wrong score of a single teacher can affect the scores of other teachers. The Houston Independent School District did not verify or audit the value-added scores which were likened to a black-box. Thus, the teachers had no meaningful way to ensure accuracy of the scores and were subject to unfair deprivation of their constitutionally protected property interests in their jobs.

At the other extreme, we should recognise there are business and organisational objections against full transparency due to the desire to maintain trade secrets and intellectual property. Here it is pertinent to highlight that full transparency is not necessarily beneficial to users and laypersons (SAL Report 2020, para. 2.50). The disclosure of source code is not always necessary and, as we will discuss below, may not be sufficient for ensuring the fairness of the AI process for users and affected persons.

A related point is the presence of trade-offs between transparency and accountability to the affected persons. Employing organisations may not want certain criteria they have utilised in selecting candidates to be disclosed. Yet the candidate who has reasonable grounds to believe he has been unfairly excluded from the job opportunity by the AI system should be in a position to legitimately demand some assurance that the AI output at hand was derived via a fair process.

To mitigate the opacity problem, the AI processes or outcomes should ideally be explainable. Explainable AI can act as a check on bias and errors, and potentially enhance trust amongst stakeholders. Additionally, it can unravel the steps or criteria used in the algorithmic process and such information might be useful for resolving disputes should they arise subsequently. An additional benefit might be the enhanced performance of AI systems in future should we obtain a better understanding of how the systems function.

Similar to the ethical basis of the equal opportunity merit principle, explainable AI is "intrinsically valuable" in its respect for human personhood and dignity (Colaner 2021). Importantly, explainable AI can aid in our quest to reap the benefits of incorporating the equal opportunity merit principle to promote fairness in AI-driven employment decision-making, and, in particular, to ensure the application of appropriate job-related factors by the employing organisation.

At its core, a proper explanation should be capable of answering the "why" question in context or not merely in an abstract fashion. With respect to AI employment decision-making, it is contended that the explanation should be able to provide an answer to the question "why did the AI generate this or that specific employment decision or recommendation?" The concept of explanation is intimately associated with the notions of interpretation and understanding. Explainable AI refers to "a human-interpretable description" of the AI process that allows the observer to "determine the extent to which a particular input was determinative or influential on the output" (Doshi-Velez and Kortz 2017). This implies that, upon the explanation being given, users should be made aware of the determinative factors for the AI-generated decisions.

It is proposed that an adequate answer to the "why" and "how" questions must be communicated in a manner capable of being understood by a reasonable user of the AI (the employing organisation) and affected persons (the employees or job candidates). The content of the explanation may include the functioning of the AI model and/or the process in arriving at the specific AI predictions or recommendations. This approach not only requires "explicability" to ensure individuals obtain an explanation of the AI decision-making processes (Floridi et al. 2018, p. 702) and understand the outcomes (OECD 2019, para. 1.3) but also adopts an objective user/person-centric perspective in assessing explainability. This further emphasises the bilateral nature of the communication process.

Take, for example, an aggrieved candidate for a job application who requests for an explanation of the specific AI decision taken against the candidate due to suspicions of bias or inaccuracies. The AI model used by the organisation for employment decisions may provide information on the determinative factors that result in the specific decision taken against the candidate. The decision not to shortlist a candidate for a job interview may, for instance, be explained by reference to the AI model that disregards all applicants without a particular qualification such as a professional degree or a particular type of prior job experience.

Footnote 9 (continued)

in only one direction, and at a magnitude represented by a readily available coefficient.".

[10] Civil Action H-14-1189.

In other cases, a more targeted explanation catered to the candidate's specific situation (e.g., by referencing a combination of factors and their relative weights vis-à-vis the particular employment decision) may be required. Whether the explanation is sufficient or not should be assessed from the viewpoint of a reasonable candidate taking into account his or her general knowledge of the employment situation, the industry and technological processes.

The General Data Protection Regulation (GDPR)[11] refers to the "right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her".[12] Where the person is subject to such automated processing, he has the right to "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" (Articles 13–15; Selbst and Powles 2017). It was argued that one must seek explanations of the process behind a model's development, not just explanations of the model itself (Selbst and Barocas 2018). Further, Robbins (2019, p. 500) focused on "meaningful human control" primarily of the user over the algorithms. Whether the information in question is "meaningful" to users and affected persons should depend on the context. For a candidate subject to a job screening process, for example, meaningful information would probably include the logic of the criteria applied to screen job candidates and the impact they have on whether he or she would be shortlisted or rejected for the job.

Explainable AI enables human understanding of the AI decision-making so that the candidate may assess whether the AI processes have been biased or unfair. Reasonable employing organisations and human resource experts who understand the AI decision-making processes and/or outcomes should be in a position to accept or alternatively, to overrule, the AI recommendation if there is suspicion of bias. At the same time, an adequate understanding of AI decision-making would assure affected candidates that they have not been subject to an unfair process. As is consistent with the equal opportunity merit principle, users and affected persons must be entitled to exercise their moral powers to seek a conception of justice and the good (in this case, fair employment opportunities) within the confines of the legal system by mounting a challenge against biased AI employment decision-making processes.

How should the appropriate standard of explanation based on the user/person-centric perspective be conceptualised and operationalised? The standard of human understanding should be based on that of an ordinary reasonable person without special expertise of AI but who possesses sufficient general knowledge including knowledge about basic logical reasoning, the general relationship between cause and effect, employment matters, human traits and behaviours. The appropriate method of explanation will depend on the perspectives and circumstances of the stakeholders. In addition to the features of a reasonable user with general knowledge, such a reasonable user would also likely prefer an explanation that fits with practical and reasonable notions of non-bias and fairness such as the equal opportunity merit principle.

In practice, the human resources personnel of the employing organisation, who may not possess any AI or data science expertise, would have to first understand and interpret the AI decision-making from AI designers and/or vendors. Based on their understanding and interpretation of the AI process, the employing organisation may have to select the relevant information to be provided to the candidates for the purpose of explanation when called upon. Mittlestadt et al. (2019) reminded us that explainable AI is not only concerned about how the recipient of an explanation perceives it but also involves communication exchange and dialogue between the giver and recipient.

We should also briefly examine the types of explanations and whether they are conducive for the reasonable user's understanding the AI output. There are various types of explanation including textual or visual explanations of the relationship between the features of the inputs and the outputs, comparing cases in the training data that are analogous to the decision at hand, and providing local explanations by explaining the fit between the model to a particular decision e.g., local interpretable model-agnostic explanations or LIME (Ribeiro et al. 2016; Yao 2021). The last example suggests that, on occasions, complex AI models may not be amenable to a full explanation except by recourse to a simpler model that approximates the actual model (Baum et al. 2022). Thus, there may be trade-offs between accuracy and explanation. Furthermore, certain types of explanation may be more appropriate to the employment context than others. This can be assessed by reference to three types of explanations namely probabilistic, counterfactual and contrastive explanations.

Probabilistic explanations are those that focus on why the AI model had a certain level of confidence associated with certain attributes or features. For example, the fact that candidate has an accounting degree increases the probability by X% that he gets the job. There may be a number of factors (e.g., qualifications, prior experience in tax, postal code address) associated

[11] Recital 71. Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC, 2016 O.J. (L 119) 1 (EU).

[12] The scope covers "any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work" amongst others.

with varying probabilities that the candidate is to be shortlisted or recommended. This can complicate the decision-making process and is not likely to be intuitive to a reasonable user.

Altman et al. (2018) suggested the use of counterfactual analysis in explaining algorithmic decisions, for example, by examining the "counterfactual causal estimation of the effects of an automated decision on the well-being of an individual". Essentially, this approach relies on the counterfactual explanation of the effects of a decision with a focus on the material difference in the outcomes arising from the automated decision. One advantage of this type of explanation is that there is no need for the user to appreciate the internal logic of the AI model (Wachter et al. 2018, p. 851). It can also aid in assessing whether an algorithmic decision is fair e.g., whether a candidate would have been shortlisted for the job if he or she had been of another race (Wachter et al. 2018, p. 853; Kusner et al. 2017). In short, counterfactual explanations can respond more directly to the specific concerns of the job candidate than probabilistic explanations.

Causal explanations may also involve an appeal to a counterfactual cause or event which did not occur. This is known as a "contrastive explanation". According to Miller (2018), when people ask for an explanation of an event, they may be asking for an explanation relative to some contrast case (i.e., "Why P rather than Q?"). Lipton (1990) defined the answer to a contrastive question as the Difference Condition i.e., "[t]o explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the history of not-Q." In this way, the contrastive question plays a role in imposing a further restriction on the possible causes (Lipton 1990, p. 250) and therefore helps to determine an explanatory cause (Lipton 1990, p. 257). Taking employment screening as an illustration, one relevant question would be: why did the AI shortlist candidate X instead of candidate Y especially if both candidates are quite close in terms of attributes? The answer may, for example, lie in an attribute which candidate X possessed but which candidate Y did not. This seems to be the sort of question that an aggrieved candidate who suspects bias would be interested to find out from the employing organisation. Significantly, the answer to be obtained would be most relevant to employing organisations and candidates who find the equal opportunity merit principle persuasive.

5 Integrating equal opportunity merit and explainable AI for employment decision-making

It is clear from the preceding sections that explainable AI is valuable and can support and reinforce the equal opportunity merit principle which conduces to fairness in employment decision-making. However, these values and approaches, important as they are, may come to nought if they cannot be usefully incorporated in the AI processes. The central design and implementation question thus concerns how the concepts of equal opportunity merit principle and explainable AI can be translated into the design of AI-driven employment decision-making processes.

Commentators have noted that important ethical values can be consciously embedded in the design of technology including AI (Poel 2020; Morley et al. 2020). We can explore the Glass-Box approach by Tubella et al. (2019) to "[map] moral values into explicit verifiable norms that constrain and direct inputs and outputs". The model is consistent with enhancing explainability of AI systems. Essentially, it seeks to ensure the AI system adheres to the designated moral values in a specific context. In the first interpretive stage, an attempt is made initially to determine the desirable values depending on the different legislative, regulatory and ethical frameworks. These abstract values are then to be translated into more "finely-grained" norms and "functionalities" relating to the inputs and outputs. At the second stage, the behaviour of the system is monitored and checked for its adherence to the more concrete requirements on inputs and outputs as determined at the first stage.

The translation process should take into account ethical as well as legal and regulatory standards. On the ethical level, we have noted that the equal opportunity merit principle is, first and foremost, part of Rawlsian ethical theory. We also find expression of the principle in ethical codes on employment. For example, the non legally-binding ethical guidelines in Singapore[13] enjoin employers to refrain from indicating a cut-off age for recruitment but instead state the specific job requirements such as the need for physical handling of heavy equipment. Furthermore, the equal opportunity merit principle underlies the employment laws and regulations at the international and national level respectively. The International Labour Organization's Discrimination (Employment and Occupation) Convention 1958 (No. 111)[14] sounded the clarion call for "equality of opportunity and treatment in respect of employment and occupation" amongst Member States. In the US, the Civil Rights Act (Title VII)[15] sought to eliminate race-based discrimination and other forms of

[13] Tripartite Guidelines on Fair Employment Practices by the Tripartite Alliance for Fair & Progressive Employment Practices at https://www.tal.sg/tafep/Getting-Started/Fair/Tripartite-Guidelines.

[14] http://www.ilo.org/dyn/normlex/en/f?p=NORMLEXPUB:12100:0::NO::P12100_ILO_CODE:C111, Articles 1 and 2. A total of 175 countries have ratified the Convention as of September 2021.

[15] Title VII of the Civil Rights Act of 1964, as amended by the Civil Rights Act of 1991, 42 U. S. C. §§2000e–2(a).

employment bias[16] based on the legal tests of intentional discrimination (ie, differential treatment at the expense of a protected group) and disparate impact respectively (Ajunwa 2020b). Equal opportunities commissions, human rights or similar commissions in the UK,[17] Hong Kong[18] and New Zealand[19] are responsible for enforcing non-discrimination laws and promoting equal opportunities in employment contexts. Certain principles e.g., that employers should omit, in job applications, certain fields requiring information on age, race, religion, and other protected characteristics are found in both ethical codes[20] and statutes.[21]

Though the two-stage approach was designed for intelligent systems generally including neural networks to agent-based systems according to Tubella et al. (2019), we can adapt it for the use of AI in the employment context, for example, in job recruitment. At the first interpretive stage, it is first necessary to determine what fairness as a normative concept entails by reference to national regulatory, legal or ethical frameworks which we have mentioned above. In the context of employment decision-making, the moral value of fairness may be explicitly concretised at the initial level in the form of the equal opportunity merit principle. The latter can be further concretised into the job-related merit factors that underlie the equal opportunity merit principle. The specific job-related merit factors for each industry or for different jobs within the same industry may vary. Nevertheless, typical factors might include relevant academic qualifications and/or professional qualifications, working and/or internship experiences, job-specific skills, language proficiency and communication skills. On the other side of the coin, as mentioned above, the exclusion of irrelevant factors from the AI process is equally important. These selected factors will also have to be aligned with the content of the ethical codes, and legal and regulatory frameworks that may specifically prohibit employment discrimination based on certain protected characteristics.

The choice of inputs (e.g., the data extracted from the candidates' CVs) for the AI system would be constrained by the selected merit factors for the job and this would in turn affect the outputs (e.g., the recommendation as to whether the candidate should be shortlisted for the interviews). The linkage between the job-related factor such as academic and/professional qualifications and the concrete functionalities relating to the inputs (e.g., the academic/professional transcript of candidate) can be quite straightforward in the standard cases. In other cases, discretion may have to be exercised by the human resource department as to whether a particular job-related factor (e.g., a job-specific skill) should be connected to a particular functionality as input.

Determining the concrete functionality or group of functionalities e.g., psychometric tests for more intangible attributes such as the desired personalities and moral virtues of the candidates can be challenging. The relevance and validity of the psychometric tests taken by candidates to predict job performance and/or fit with the employing organisation should be empirically tested with respect to factors such as job type, tenure and work attitudes (Sekiguchi and Huber 2011; Arthur et al. 2006). In addition, the AI tools that are utilised in conjunction with the chosen job-related factors and functionalities should be informed by empirical evidence as to their relevance, validity and reliability in predicting the desired outcomes in employment (Tippins et al. 2021).

AI can assist in making the assessment of job-related merit factors more objective and consistent. A chat-based structured interview tool for conducting online interviews utilised natural language processing and machine learning to develop a regression model that generated inferences from the textual answers regarding the personality traits of candidates which were in turn validated by 117 volunteers with an accuracy of 87.83% (Jayaratne and Jayatilleke 2020). AI-based video interviews were perceived by job applicants to be fairer procedurally and more consistent and objective than traditional evaluation procedures e.g., where the algorithm was designed to ensure parity in the questions asked of job applicants in terms of the level of difficulty and the response times for candidates' responses regardless of their appearances or other characteristics (though concerns were expressed about potential data bias and lack of diversity of candidates) (Kim & Heo 2022).

The fairness algorithm to be selected for the AI system should also be aligned with the equal opportunity merit principle. Taking into consideration the different conceptions of AI fairness, bias mitigation algorithms such as Google's "What if" tool[22] allow the employing organisation to visualise the effects of different bias mitigation strategies and metrics in order to determine which fairness metric to use.

[16] See *Griggs v. Duke Power Co*. 401 US 424 (1971); and *United Steelworkers of America v Weber* 443 US 193 (1979).

[17] The UK Equality and Human Rights Commission promotes equal opportunities at the workplace under the Equality Act 2010: https://www.eoc.org.uk/.

[18] https://www.eoc.org.hk/en/about-the-eoc/introduction-to-eoc.

[19] The Human Rights Commission under the NZ Human Rights Act 1993 at https://www.hrc.co.nz/about/vision-mission-values-and-statutory-responsibilities/. See also the Employment Relations Act 2000.

[20] See Singapore's Tripartite Guidelines on Fair Employment Practices by the Tripartite Alliance for Fair & Progressive Employment Practices at https://www.tal.sg/tafep/Getting-Started/Fair/Tripartite-Guidelines.

[21] UK Equality Act 2010.

[22] https://pair-code.github.io/what-if-tool/.

Specific to the employment context, the algorithm should ideally allow for assessments of individual fairness in order to make decisions on the suitable candidate amongst two or more candidates with apparently similar attributes. In this regard, as mentioned above, certain explanations of AI output based on counterfactual and/or contrastive explanations would be more appropriate.

Finally, at the second stage, it is suggested that the concrete norms (i.e., the job-related factors) be machine-verifiable and that the system allow for expeditious monitoring of compliance between these factors and the "functionalities" (e.g., the specific data from CVs) (Tubella et al. 2019). Legal requirements relating to AI audits to be conducted by independent auditors (e.g., Hilliard et al. 2022 on New York City's new law requiring bias audits of AI-driven employment decision-making) should be noted.

6 Conclusion

Bias can potentially infect the AI employment decision-making process at multiple stages, and AI opacity presents a further obstacle to the attainment of fairness of the AI process and output. In this regard, the equal opportunity merit principle based on Rawls' theory with its emphasis on merit job-based factors offers an ethical approach for employing organisations to ensure that the various stakeholders' interests are taken into account in an open and objective manner. Given the limitations of attaining full transparency of the AI model for employment decision-making, employing organisations should focus on explainable AI that allows its outputs to be reasonably understood by the human resource practitioners, employees and job candidates in the specific context of employment. Arguably, a person-centric perspective to explainable AI that incorporates counterfactual and/or contrastive explanations is more conducive for employment decision-making. Significantly, explainable AI supports the underlying aim of the equal opportunity merit principle, and as such, both are well-aligned with the goal of fairness in AI employment decision-making. Some preliminary ideas have been explored with a view to integrating these two important ethical approaches in a proposed value-based AI design to implement fair employment practices.

Acknowledgements This research is supported by the National Research Foundation, Singapore under its Emerging Areas Research Projects (EARP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author's and do not reflect the views of National Research Foundation, Singapore.

Declarations

Conflict of interest There is no conflict of interests.

References

Altman M, Wood A, Vayena E (2018) A harm-reduction framework for algorithmic fairness. IEEE Secur Priv 16(3):34–45

Ajunwa I (2020a) The "black box" at work. Big Data Soc 7(2):1–6

Ajunwa I (2020b) The paradox of automation as anti-bias intervention, 41 Cardozo L Rev 1671

Arneson RJ (1989) Equality and equal opportunity for welfare". Philos Stud 56(1):77–93

Arthur W, Bell ST, Villado AJ, Doverspike D (2006) The use of person organization fit in employment decision making: an assessment of its criterion-related validity. J Appl Psychol 91(4):786–801

Barocas S, Selbst A (2016) Big data's disparate impact. Calif Law Rev 104(3):671–732

Baum K, Mantel S, Schmidt E, Speith T (2022) From Responsibility to reason-giving explainable artificial intelligence. Philos Technol 35:12

Bellamy RKE, et al (2018) AI Fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. https://arxiv.org/abs/1810.01943

Binns R (2018) Fairness in machine learning: lessons from political philosophy. Proc Mach Learn Res 81:1–11

Bogen M, Rieke A (2018) Help wanted: an examination of hiring algorithms, equity and bias. https://www.upturn.org/static/reports/2018/hiring-algorithms/files/Upturn%20--%20Help%20Wanted%20-%20An%20Exploration%20of%20Hiring%20Algorithms,%20Equity%20and%20Bias.pdf. Accessed 16 June 2022

Burrell J (2016) How the machine "thinks": understanding opacity in machine learning algorithms. Big Data Soc 3:1

Calders, T & Zliobaite, I (2013) Why unbiased computational processes can lead to discriminative decision procedures. In: Discrimination and privacy in the information society (Vol 3, pp 43–57). (Studies in Applied Philosophy, Epistemology and Rational Ethics). Springer. https://doi.org/10.1007/978-3-642-30487-3_3

Chalfin A, Danieli O, Hillis A, Jelveh Z, Luca M, Ludwig J, Mullainathan S (2016) Productivity and selection of human capital with machine learning. Am Econ Rev 106(5):124–127

Chamorro-Premuzic T, Akhtar R (2019) Should companies use AI to assess job candidates? https://hbr.org/2019/05/should-companies-use-ai-to-assess-job-candidates. Accessed 16 June 2022

Cohen GA (2011) On the currency of egalitarian justice and other essays in Political Philosophy. Princeton University Press

Colaner N (2021) Is explainable artifcial intelligence intrinsically valuable? AI & Soc. https://doi.org/10.1007/s00146-021-01184-2

Corbett-Davies, S and Goel, S (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. https://arxiv.org/abs/1808.00023

Davis JL, Williams A, Yang MW (2021) Algorithmic reparation. Big Data Soc 8(2):1–12

Doshi-Velez, F., and Kortz, M. (2017). Accountability of AI under the law: the role of explanation. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper. https://arxiv.org/abs/1711.01134

Dwork C Hardt M, Pitassi T, Reingold O, Zeme RS (2012) Fairness through awareness. Proceedings in 3rd Innovations in Theoretical Computer Science. Cambridge, MA, USA, January 8–10, 214–226

Dworkin R (2000) Sovereign Virtue: the theory and practice of equality. Harvard University Press, Cambridge

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

Friedman B, Nissenbaum H (1996) Bias in computer systems. ACM Transact Inform Syst 14(3):330–347
Greenwald AG, Banaji MR (1995) Implicit social cognition: attitudes, self-esteem, and stereotypes. Psychol Rev 102(1):4–27
Greenwald A, Krieger LH (2006) Implicit bias: scientific foundations. Calif Law Rev 94(4):945–967
Hall P, Gill N (2018) An introduction to machine learning interpretability. Sebastopol, CA: O'Reilly Media
Harrison DA, Kravitz DA, Mayer DM, Leslie LM, Lev-Arey D (2006) Understanding attitudes toward affirmative action programs in employment: summary and meta-analysis of 35 years of research. J Appl Psychol 91(5):1013–1036
Heinrichs B (2021) Discrimination in the age of artificial intelligence. AI Soc. https://doi.org/10.1007/s00146-021-01192-2
Hilliard A, Kazim E, Koshiyama A, Zannone S, Trengove M, Kingsman N, Polle R (2022) Regulating the robots: NYC mandates bias audits for Ai-driven employment decisions (April 13, 2022). Available at SSRN: https://ssrn.com/abstract=4083189 or https://doi.org/10.2139/ssrn.4083189. Accessed 16 June 2022
Holmes E (2005) Anti-discrimination rights without equality. Mod Law Rev 68(2):175–194
Houser KA (2019) Can AI solve the diversity problem in the tech industry: mitigating noise and bias in employment decision-making. Stanford Technol Law Rev 22:290
Jayaratne M, Jayatilleke B (2020) Predicting personality using answers to open-ended interview questions. IEEE Access 8:115345–115355. 10. 1109/ACCESS.2020.3004002
Kim J-Y, Heo WG (2022) Artificial intelligence video interviewing for employment: perspectives from applicants, companies, developer and academicians. Inf Technol People 35(3):861–878
Kroll JA, Huey J, Barocas S, Felten EW, Reidenberg JR, Robinson DG, Yu H (2017) Accountable algorithms. Univ Pa Law Rev 165:633–707
Kusner MJ, Loftus JR, Russell C et al (2017) Counterfactual fairness. https://arxiv.org/abs/1703.06856
Lee MSA, Floridi L, Singh J (2021) Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. AI Ethics. https://doi.org/10.1007/s43681-021-00067-y
Lipton P (1990) Contrastive explanation. R Inst Philos Suppl 27:247–266
Miller T (2018) Contrastive explanation: a structural-model approach. https://arxiv.org/abs/1811.03163
Mittelstadt B, Russell C, Wachter S (2019) Explaining explanations in AI. In FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, 2019, Atlanta, GA, USA. ACM, New York, NY, USA. https://doi.org/10.1145/3287560.3287574
Morley J, Floridi L, Kinsey L, Elhalal A (2020) From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 26(4):2141–2168. https://doi.org/10.1007/s11948-019-00165-5
OECD (2019) Recommendation of the Council on Artificial Intelligence. Retrieved from https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Accessed 1 June 2022
Poel I (2020) Embedding values in artificial intelligence (AI) systems. Mind Mach 30(3):385–409. https://doi.org/10.1007/s11023-020-09537-4
Raghavan M, Barocas S, Kleinberg J, Levy K (2019) Mitigating bias in algorithmic employment screening: evaluating claims and practices. https://arxiv.org/pdf/1906.09208.pdf
Rawls J (1971) A theory of justice. Oxford University Press
Rawls J (1999) The law of peoples. Harvard University Press
Rawls J (2001) Justice as fairness: a restatement. The Belknap Press of Harvard University Press
Ribeiro MT, Singh S, Guestrin C (2016) Why Should I Trust You? Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, 1135–1144
Robbins S (2019) A misdirected principle with a catch: explicability for AI. Mind Mach 29:495–514
Roemer J (2000) Equality of opportunity. Harvard University Press
Romei A, Ruggieri S (2014) A multidisciplinary survey on discrimination analysis. Knowledge Eng Rev 29(5):582–638
Ryan M (2020) In AI we trust: ethics, artificial intelligence, and reliability. Sci Eng Ethics 26:2749–2767. https://doi.org/10.1007/s11948-020-00228-y
Sandel MJ (2021) The Tyranny of Merit—What's Become of the Common Good? Penguin Random House UK
Selbst AD, Barocas S (2018) The intuitive appeal of explainable machines. Fordham Law Rev 87:1085
Selbst AD, Powles J (2017) Meaningful information and the right to explanation. Int Data Privacy Law 7(4):233–242
Sen A (1992) Inequality examined. Harvard University Press, Cambridge Massachusetts
Sekiguchi T, Huber VL (2011) The use of person–organization fit and person–job fit information in making selection decisions. Organ Behav Hum Decis Process 116:203–216
Sinclair A, Carlsson R (2021) Reactions to affirmative action policies in hiring: Effects of framing and beneficiary gender. Anal Soc Issues Public Policy 21:660–678
Singapore Academy of Law (SAL) (Law Reform Committee), subcommittee on Robotics and Artificial Intelligence. (2020). Applying Ethical Principles for Artificial Intelligence in Regulatory Reform
Tambe P, Cappelli P, Yakubovich V (2019) Artificial intelligence in human resources management: challenges and a path forward. Calif Manage Rev 61(4):15–42
Temkin LS (2016) The many faces of equal opportunity. Theory Res Educ 14(3):255–276
Tippins N, Oswald F, McPhail SM (2021) Scientific, legal, and ethical concerns about AI-based personnel selection tools: a call to action. Personnel Assessment Decisions. https://doi.org/10.25035/pad.2021.02.001
Tubella AA, Theodorou A, Dignum F, Dignum V (2019) Governance by glass-box: implementing transparent moral bounds for AI behaviour. https://arxiv.org/abs/1905.04994
Wachter S, Mittelstadt B, Russell C (2018) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv J Law Technol 31:841
Yao YH (2021) Explanatory pluralism in explainable AI. https://arxiv.org/abs/2106.13976