Edukacja Filozoficzna 79/2025
ISSN 0860-3839, eISSN 2956-8269
DOI: 10.14394/edufil.2025.0004
ORCID: 0009-0001-8886-6620

Justice and AI Fairness: John Rawls and Iris Marion Young on Racist and Sexist AI Decisions

Neomal Silva
(Independent researcher and Senior ICT Consultant, Melbourne, Australia)

Abstract: AI outcomes that exhibit racism, sexism, homophobia, or other biases are deemed "unfair." Several scholars have applied John Rawls's theory of justice to evaluate this unfairness. This paper clarifies, though, that Rawls's ideal and nonideal theories are ill-equipped to deal with individual instances of AI unfairness; it furthermore argues that Young's theory is better equipped to do so – not only because it includes sociological accounts of racism and other -isms, but also because it incorporates the consciousness-raising spaces that help "name" the racist, sexist, etc. behaviours – behaviours that, if left unnamed, remain undetected, and, as a result, are both re-enacted in society and reproduced by AI.

Key words: AI bias, AI fairness, Rawls, structural power, Iris Marion Young

1. Introduction

Whilst artificial intelligence (AI) encompasses robotics, rule-based systems, machine learning, and other technologies, it is machine learning, in particular, that has provided several instances of bias against marginalized groups – such as non-white people and females. Consider the following practical instances of AI bias.[1]

Amazon's recruitment team used an algorithm to rate CVs from one to five stars, only to find it favoured male candidates. The bias stemmed from training

[1] I use the terms "fairness" and "bias" (or "AI fairness" and "AI bias") interchangeably: "AI fairness" aims to ensure that no group – defined by some socially salient trait like gender or ethnicity – is unfairly disadvantaged. "AI bias," on the other hand, refers to the unfair or skewed outcomes that discriminate against certain groups. In essence, "AI fairness" is the goal, whilst "AI bias" refers to the obstacles to achieving it. For an overview of definitions of different types of AI fairness and AI bias, along with a survey of different data-centric techniques for mitigating bias, see E. Ferrara, *Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies*, "Sci" 2024, Vol. 6, No. 1, pp. 1–15, https://doi.org/10.3390/sci6010003.

the algorithm on a dataset made up of CVs of people previously hired for the role – most of whom were men.[2]

Joy Buolamwini discovered that commercial facial recognition technologies from companies like IBM, Microsoft, and Megvii had higher error rates for darker-skinned people and women.[3] The bias arose because the algorithms were primarily trained on faces of young white men.

Google launched a photos app designed to categorize users' photos but faced backlash when it miscategorized African Americans as "gorillas." This offensive error occurred because the algorithm lacked sufficiently diverse training data.

To examine instances of AI *unfairness* such as these, scholars might turn to John Rawls's concept of justice as *fairness*. Whilst some have used Rawls's work to study AI ethics,[4] Morten Bay cautions against oversimplifying or taking Rawls's ideas out of context.[5] Nonetheless, scholars have engaged with Rawls in their studies of AI bias and fairness.[6] For example, Flavia Barsotti and Rüya Gökhan Koçer argue that Rawls's *Theory of Justice* "provides the foundations to a solution

[2] J. Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women*, Reuters, 9.10.2018, URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[3] J. Buolamwini, T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, "Proceedings of Machine Learning Research" 2018, Vol. 81, p. 8.

[4] E.g., I. Gabriel, *Toward a Theory of Justice for Artificial Intelligence*, "Daedalus" 2022, Vol. 151, No. 2, pp. 218–231, https://doi.org/10.1162/daed_a_01911; H. Heidari et al., *Fairness behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making*, "Advances in Neural Information Processing Systems" 2018, Vol. 31; R. Binns, *Algorithmic Accountability and Public Reason*, "Philosophy and Technology" 2018, Vol. 31, p. 543; L. Weidinger et al., *Using the Veil of Ignorance to Align AI Systems with Principles of Justice*, "Proceedings of the National Academy of Sciences of the United States of America" 2023, Vol. 120, e2213709120, https://doi.org/10.1073/pnas.2213709120.

[5] M. Bay, *Participation, Prediction, and Publicity: Avoiding the Pitfalls of Applying Rawlsian Ethics to AI*, "AI and Ethics" 2024, Vol. 4, p. 1545, https://doi.org/10.1007/s43681-023-00341-1.

[6] E.g., see F. Barsotti, R.G. Koçer, *MinMax Fairness: From Rawlsian Theory of Justice to Solution for Algorithmic Bias*, "AI & Society" 2024, Vol. 39, pp. 961–974, https://doi.org/10.1007/s00146-022-01577-x; A.K. Jørgensen, A. Søgaard, *Rawlsian AI Fairness Loopholes*, "AI and Ethics" 2022, Vol. 3, pp. 1185–1192, https://doi.org/10.1007/s43681-022-00226-9; T. Krupiy, *A Vulnerability Analysis: Theorising the Impact of Artificial Intelligence Decision-Making Processes on Individuals, Society and Human Diversity from a Social Justice Perspective*, "Computer Law & Security Review" 2020, Vol. 38, 105429, https://doi.org/10.1016/j.clsr.2020.105429; L.M. Rafanelli, *Justice, Injustice, and Artificial Intelligence: Lessons from Political Theory and Philosophy*, "Big Data and Society" 2022, Vol. 9, No. 1, https://doi.org/10.1177/20539517221080676.

for algorithmic bias"[7] – where the algorithmic bias is against "gender, ethnicity, disability, etc."[8] To offer another example: Anna Katrine Jørgensen and Anders Søgaard, though they critique the use of Rawls to achieve algorithmic fairness, assume his difference principle can be applied to "groups […] typically thought of as the product of a subset of protected attributes, e.g., gender and race."[9] However, Rawls's difference principle[10] is concerned with income groups, not groups defined by protected attributes. In Rawls's framework, the "worst off" refers to those with the least income or wealth, and economic inequality is allowed only if it benefits the absolute position of that socioeconomically disadvantaged group. It should be apparent, then, that scholars should tread carefully when applying Rawls's ideas to AI fairness.

One aim of this paper is not only to urge AI fairness scholars to exercise caution when applying Rawlsian concepts, like the difference principle or the veil of ignorance, but also to argue a stronger claim: fundamentally, Rawls's theory is ill-equipped to address biases related to race, gender, and other forms of discrimination in AI. This is partly because Rawls abstracts from structural power – a type of power implicated in racism, sexism, and other -isms[11] – but also because his ideal and nonideal theories are not designed to tackle specific instance of social injustice (like biased machine-learning outputs). Though A. John Simmons[12] has

[7] F. Barsotti, R.G. Koçer, *MinMax Fairness*, op. cit., p. 961. It is also too big a jump to go from Rawls's *Theory of Justice* – which concerns the two principles of justice that Rawls argues should govern the basic structure of society (e.g., the constitution) – to immediately proposing that Rawls's two principles ought to constrain the outputs of a machine-learning algorithm. Iason Gabriel, in his *Toward a Theory of Justice for Artificial Intelligence*, points out that technology (and AI, in particular) cannot be assumed to be part of the basic structure – i.e., it cannot be assumed to be the part of the subject of Rawls's two principles of justice – but Gabriel argues strongly for its inclusion. See I. Gabriel, *Toward a Theory of Justice*, op. cit.

[8] F. Barsotti, R.G. Koçer, *MinMax Fairness*, op. cit., p. 964.

[9] A.K. Jørgensen, A. Søgaard, *Rawlsian AI Fairness Loopholes*, op. cit., p. 1187.

[10] J. Rawls, *A Theory of Justice*, Harvard University Press, Cambridge, MA, 1971, p. 83.

[11] I define racism and sexism much like Iris Young understands them. She views "racism" as a systemic and structural phenomenon that marginalizes and disadvantages racial groups. This occurs through institutional practices, cultural norms, and social policies that perpetuate racial inequalities. Racism, in this sense, goes beyond overt discrimination or prejudice and includes the ways societal institutions maintain and reproduce racial hierarchies. Similarly, "sexism," in Young's view, is a structural form of oppression that subordinates women and reinforces gender roles through societal norms, institutions, and practices. Not limited to individual acts of discrimination, it furthermore encompasses the pervasive behavioural norms that perpetuate gender inequality and limit women's opportunities.

[12] A.J. Simmons, *Ideal and Nonideal Theory*, "Philosophy & Public Affairs" 2010, Vol. 38, pp. 5–36.

argued that Rawls's theories are not suited to addressing specific social injustices outside the context of AI, this critique is yet to be articulated in AI fairness literature. I will articulate it here, as it is vital to prevent scholars from misapplying Rawls's theories to challenges his work is not equipped to solve.

A second aim of this paper is to propose Iris Marion Young's critical theory of social justice as an alternative to Rawls's theory. Unlike Rawls's, Young's theory is deeply connected to sociological accounts of structural power. I will show that engagement with structural power is essential for evaluating unfairness in AI decision-making, making Young's theory the preferable approach. Crucially, her theory provides the conceptual tools to expose the very -isms that are reproduced in the AI outcomes that draw the most media criticism – such as gender-biased recruitment,[13] racist image classification,[14] antisemitic messaging,[15] and over-policing of certain ethnicities.[16]

I proceed as follows. In section 2, I provide Rawls's accounts of what is "just," what is "unjust," and what is "permissible," and I clarify that these accounts are not intended to deal with single instances of unfairness. Notably, none of Rawls's accounts (of what is "just," "unjust," etc.) refer to structural power. In section 3, we consider structural power, using an example to illuminate some of its complexities, along with some of the consequences it can have for those disadvantaged by it. That elucidation helps confirm that Rawls's theory is not equipped to attend to the kinds of injustices that worry AI ethicists. Its disregard for structural power may prompt philosophers to seek a theory that does engage with it. In section 4, we turn to one such theory – Young's feminist critical theory. We note its ability to capture the power that resides at what Anthony Giddens calls the level of "practical consciousness." Moreover, we examine its engagement with discursive consciousness-raising spaces – that is, the spaces in which structural

[13] Reuters, *Amazon Ditched AI Recruiting Tool that Favored Men for Technical Jobs*, "The Guardian," 11.10.2018, URL: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine.

[14] M. Zhang, *Google Photos Tags Two African-Americans as Gorillas through Facial Recognition Software*, "Forbes," 1.07.2015, URL: https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/.

[15] S. Buranyi, *Rise of the Racist Robots: How AI Is Learning All Our Worst Impulses*, "The Guardian," 8.08.2017, URL: https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses.

[16] CBC Radio, *Police Are Considering the Ethics of AI, Too*, 21.09.2018, URL: https://www.cbc.ca/radio/spark/tech-in-policing-1.4833189/police-are-considering-the-ethics-of-ai-too-1.4833194.

oppression has historically found its voice. The attributes of Young's critical theory not only enable us to conceptualize structural power but also equip us with the tool – namely, consciousness-raising spaces – that could help liberate society from its various -isms. In section 5, we investigate the implications of these insights for AI decision outcomes. We also address potential objections. Section 6 offers concluding remarks.

2. Rawls's Ideal and Nonideal Theories of Justice

Rawls's theory of "justice as fairness" argues that a just society (i) institutes socio-economic inequalities only if they benefit the "worst off," and (ii) ensures all members have equal basic liberties and fair equality of opportunity. This is achieved through abstractions like the "original position"[17] and "veil of ignorance,"[18] where rational individuals would choose these two principles of justice when they are unaware of their own social status or of their own personal characteristics and talents.

Some philosophers argue that Rawls's theory is ill-equipped to address issues like sexism, racism, and other -isms. It "abstracts from the determinate content of social life,"[19] they say, ignores "the importance of social groups,"[20] and is mute on how "to rectify [racial] injustices that have already occurred."[21] Rawls, of course, offers us an ideal theory of justice (as outlined above) – but also a nonideal theory. His ideal theory elucidates an abstract conception of justice, whilst his nonideal theory articulates how to move us closer to it – without that nonideal theory necessarily attempting to eliminate particular instances of injustice, such as -isms. This requires some explanation.

In his ideal theory, Rawls articulates the constitutional principles that citizens would choose from behind a veil of ignorance, that is, choose under conditions where potential biases influencing their judgment are hidden from view. Rawls

[17] J. Rawls, *A Theory of Justice*, op. cit., pp. 118–194.

[18] Ibid., pp. 136–141.

[19] L. McNay, *Recognition as Fact and Norm: The Method of Critique*, in: *Political Theory: Methods and Approaches*, eds. D. Leopold, M. Stears, Oxford University Press, Oxford 2008, p. 87.

[20] I.M. Young, *Justice and the Politics of Difference*, Princeton University Press, Princeton 1990, p. 27.

[21] See C.W. Mills, *Retrieving Rawls for Racial Justice? A Critique of Tommie Shelby*, "Critical Philosophy of Race" 2013, Vol. 1, No. 1, p. 2 (italics removed).

says that the principles so derived are "just" and that they represent the ideal to which a society ought to strive if it is to be said to be a "just" society.

Simmons, in his essay *Ideal and Nonideal Theory*,[22] responds to the complaint that Rawls's nonideal theory is silent on real-world problems, such as historical slavery,[23] and resource scarcity[24] – and his response is: it's not meant to speak to such problems. Rawls's nonideal theory does not concern itself with removing single instances of injustice per se – where such instances might include crime, or an -ism. Its purpose, instead, is to do what is required to move society from less-than-just to (Rawls's notion of) "just," as long as the actions that are taken to carry out that move are "morally permissible," "politically feasible," and likely to succeed.[25] Simmons acknowledges that Rawls is vague on those three conditions.[26] What matters for present purposes, though, is that Rawls's nonideal theory endorses attending to an -ism only if doing so moves us closer to his ideal. Indeed, non-intervention, or even introducing a new -ism, is permissible, if it is thought to be the necessary transitional path for a society to ultimately achieve (Rawls's) "just" state.[27]

We can now say the following about Rawls's framework. A society is "just" if it has fully realized his two principles of justice. It is "unjust" if it hasn't. It is "permissible" to not intervene to address an -ism.

Furthermore, assessments of what is "just," "unjust," or "permissible" can be made without considering structural power, or engaging with discourses about lived experiences of it. I contend that this omission is problematic (at least for our present purposes of considering racist etc. outcomes). I am not alone in contending this.[28] We will consider an example in which structural power is in play – not

[22] A.J. Simmons, *Ideal and Nonideal Theory*, op. cit., p. 19.

[23] C.W. Mills, *"Ideal Theory" as Ideology*, "Hypatia" 2005, Vol. 20, No. 3, p. 168.

[24] C. Farrelly, *Justice in Ideal Theory: A Refutation*, "Political Studies" 2007, Vol. 55, p. 853.

[25] J. Rawls, *The Law of Peoples*, Harvard University Press, Cambridge, MA, 1999, p. 89.

[26] A.J. Simmons, *Ideal and Nonideal Theory*, op. cit., p. 19.

[27] For support for this interpretation of Rawls's view, and an elaboration of it, see ibid., p. 23.

[28] See I.M. Young, *Structure as the Subject of Justice*, in: I.M. Young, *Responsibility for Justice*, Oxford University Press, Oxford 2011, https://doi.org/10.1093/acprof:oso/9780195392388.003.0002, where she argues that structural power is the subject of justice, and that *pace* Rawls his basic structure in his conception of the Just ought to factor it in. Also see L. McNay, *Recognition as Fact and Norm*, op. cit., pp. 85–105, where the author offers a critique of the kind of idealized normative reasoning we find in Rawls's theory in the first section, and in the latter part of her paper she challenges Jürgen Habermas to attend more carefully to the effects of structural power in his communicative ethics.

just to highlight the problem of omitting it, but also because the example helps convey what the complex phenomenon of structural power looks like, along with some of the impacts that it has – impacts which, I hope to show, cannot *pace* Rawls be assumed to have nothing to do with an assessment of what is just, unjust, or morally permissible in our existing social arrangements.

3. Structural Power: An Illustrative Example

In the days that followed Martin Luther King's assassination, Jane Elliot, a third-grade schoolteacher in a small rural town in Iowa, exasperated by the persistent cycles of racism within America, felt that she needed to help her classroom students understand racism in a more meaningful way. She had spoken to them about discrimination in the past. But now she wanted them to sense the anguish of the racially discriminated Other, to feel their despair, "to walk in […] [their] moccasins", as she put it.[29]

Elliot divided the students into two categories based on their eye colour. She then announced, "Blue-eyed people are better than brown-eyed people. They are cleaner than brown-eyed people. They are more civilized than brown-eyed people. And they are smarter than brown-eyed people."[30] The blue-eyed children, she added, are to receive an extra five minutes to play at lunchtime, whereas brown-eyed children are barred from playing on the playground equipment from hereon in.

Suppose that Ms Elliot furthermore segregates the classroom, confining the brown-eyed children to the back left corner, and only allowing blue-eyed children to sit at the front. In the days and weeks that follow, Suzy, a particularly intelligent (brown-eyed) student never seems to get seen by Ms Elliot when she raises her hand, perhaps because Ms Elliot has grown accustomed to not looking towards that section of the room when she asks a question.[31]

The above example allows us to provide an initial outline of what structural power looks like. To be clear, brown-eyedness (and blue-eyedness) goes beyond mere "colour" here – it's not about a relationship between one's iris and the sur-

[29] PBS Frontline, *A Class Divided*, "CosmoLearning" 1985.

[30] Ibid.

[31] This is analogous to what happened at Amazon when female job applicants (who can be said to have been "putting up their hand for a job opportunity") were screened out by the AI used by Amazon for recruitment purposes. See Reuters, *Amazon Ditched AI Recruiting Tool*, op. cit.

rounding light. Rather, at least in part, brown-eyedness acquires social significance within this classroom context in relation to blue-eyedness – that is, brown-eyedness is *not* blue-eyedness. Each of these social categories emerges as a social construct intricately interwoven with the discourses generated, perpetuated, compounded, and sometimes contested, by the students, and of course, their teacher. There is a dialectical interchange between the social categories and classroom power dynamics themselves: the categories are created by power dynamics (primarily constructed and imposed, as they were, by the teacher herself), and the categories themselves reinforce and exacerbate those power dynamics (by structuring the teacher–student and student–student interactions). The concept of power between social categories, such as "blue-eyedness" and "brown-eyedness," plays a significant role in understanding the advantages or disadvantages that members of those social categories encounter – not just the possibility of using the playground equipment, but, as Suzy finds, the power to be seen, heard, respected, and listened to as an equal.[32]

The classroom with its eye-ism is analogous to actual societies riddled with the structural power of various -isms.[33] Rawls's theory remains unswayed by such power, though. The above situation is "unjust" on his account. However, that is not due to the existence of eye-ism – but to the non-realization of Rawls's two principles of justice. Furthermore, as Simmons argues in his reading of Rawls, it would be "impermissible" to remove eye-ism if that resulted in Raymond rebelling against its removal by rallying his blue-eyed compatriots to beat up the brown-eyes and strip them further of basic liberties.

There are two messages that one can take from this. There are many non-political-theorists and activists[34] who study AI bias, and our first message is for them. Already troubled by racist image classification, sexist CV filtering, etc. – they might now also be exasperated to learn that Rawls's theory would not judge

[32] As Lois McNay points out in her critique of Habermas's ideal speech situation, power dynamics permeate interpersonal exchanges, existing before them and continuing throughout. See L. McNay, *Recognition as Fact and Norm*, op. cit., pp. 85–105.

[33] We will treat the "classroom" as though it is a "state" as we work through our reasoning – since Rawls's theory of justice applies to states (rather than classrooms).

[34] In referring to "activists," I have in mind scholars like Joy Buolamwini (Founder of the Algorithmic Justice League) – who self-identifies as an activist – but also researchers like Timnit Gebru (co-founder of Black in AI), Deborah Raji, and Safiya Noble (who says in her book *Algorithms of Oppression* that she hopes to end social injustice and change the perception of marginalized people in technology).

any of those AI outcomes to be "impermissible" in and of themselves. Our analysis hopefully makes clear that Rawls's theory is ill-suited to realize their aims. His theory is fit-for-purpose if one's purpose is to clarify what (in Rawls's view) the most perfectly just society looks like. However, it is not the correct tool if your task is to eliminate particular injustices (such as those that arise in AI decision-making). The second message is to philosophers, concerned that Rawls's framework ignores structural power if it is called upon to determine the permissibility of AI outcomes. This does not, of course, mean that structural power can be assumed to have an impact on its moral permissibility – only that it perhaps should not be ignored from the outset. For that reason, they may wish to turn to critical theory, which can consider, and critically analyse, power, when it decides on the moral permissibility of AI outcomes.

4. Young's Feminist Critical Theory

A critical theoretical approach, such as that of Iris Marion Young, is dialectically linked to sociological analysis. An assessment of the social injustice of an interpersonal arrangement, she maintains, demands a social theory about the structural power within it. Young relies on Anthony Giddens's theory of structuration,[35] as well as Pierre Bourdieu's concept of habitus, to theorize -isms, making normative recommendations on its basis – rather than in the abstract.

A thorough account of her interpretation and fusion of those two social theories can be found in her essay *Structure as the Subject of Justice*.[36] People in a social setting follow certain "rules" of engagement, many of which are implicit, but for which one risks sanction if violated; for example, queue jumping; or not saying "please" when asking a favour. When people's following of such rules is implicit, it can be said to take place at the level of "practical consciousness" – meaning the actor performs the action, without being able unambiguously to explain its logic. Furthermore, within a social setting, a person has what Giddens calls "resources" – understood (at the societal level) as both the material items one relies upon to create and produce physical goods and technologies, and the nonmate-

[35] A. Giddens, *The Constitution of Society: Outline of the Theory of Structuration*, Polity, Cambridge 1986.

[36] I.M. Young, *Structure as the Subject of Justice*, op. cit.

rial social skills that bolster a person's social power (where the latter skills could include gravitas, and the ability to persuade or manipulate others).[37] Those people in a social setting who understand its rules, and possess more resources, can be said to more powerful than those who don't.

Across her body of work, Young seeks to address the concerns of social groups within contemporary American society.[38] A "social group" is not a mere collection of individuals. It is a socially salient category that structures relations between "those to whom the category attaches" and "other people within the social setting" – relations that can be described in terms such as discrimination, stereotyping, stigmatization, exclusion, socioeconomic disadvantage, and other forms of disadvantage. The social groups that focus Young's critical theory of contemporary American society include "Blacks, Latinos, American Indians, poor people, lesbians, old people, […] the disabled"[39] and, of course, women. Throughout her *Justice and the Politics of Difference*, Young argues that such citizens tend to possess fewer resources, and find themselves in social settings in which they are less adept at following the settings' rules than the dominant group. In other words, they are less powerful due to their social group membership. I do not find this claim controversial. There are many examples of such power differentials, including those that tie to perceived rule violation by the Other: as Mary Hawkesworth notes, the implicit "rules" of discourse for members of parliament in Britain, Canada, and Australia can be characterized as "loud, aggressive, and combative" and can include "screaming, shouting, and sneering that can create no-win situations for women members. Women who adopt this combative style are ridiculed and patronized by their male counterparts, whereas women who

[37] I use the word "social power" here in Keith Dowding's sense, as that is the kind of power Young seems to be referring to, when she speaks of "power over others by means of mobilizing threats of sanction or offers of desired goods"; see I.M. Young, *Structure as the Subject of Justice*, op. cit., p. 61. Dowding's concept of "social power" includes the ability not just to threaten but to persuade A, such that A changes their preference structure to bring about an end that is different to that of A's initial preference structure. See K. Dowding, *Encyclopedia of Power*, SAGE, Thousand Oaks 2011, pp. 616–619.

[38] In the opening paragraph of *Justice and the Politics of Difference*, she declares social groups as the focus of her philosophical inquiry and then in *Equality of Whom? Social Groups and Judgments of Injustice*, she challenges the assumption "that the units we should be comparing when we make judgments of inequality are individuals"; see I.M. Young, *Equality of Whom? Social Groups and Judgments of Injustice*, "The Journal of Political Philosophy" 2001, Vol. 9, No. 1, pp. 1–18; and I.M. Young, *Justice and the Politics of Difference*, Princeton University Press, Princeton 1990, p. 3.

[39] I.M. Young, *Justice and the Politics of Difference*, op. cit., p. 14.

opt for a more demure, consultative, and collaborative style are labelled 'weak' or 'unfit' for the job."[40]

In her earlier work, *Justice and the Politics of Difference*, Young argues that sexism and other -isms occur at the level of practical consciousness – in the aversive (perhaps unintended) reactions one might have to the Other, including sexist acts,[41] homophobia,[42] ageism and ableism,[43] and racism.[44] Insofar as Giddens's notion of practical consciousness is tied to unverbalizable rule-following, I take Young to mean that these aversive sexist (and so on) reactions are themselves the silent enactment of certain "group-focused routines."[45] This is what can be understood when she says that racism etc. is "enacted in [US] society […] in informal, often unnoticed and unreflective speech, bodily reactions to others, conventional practices of everyday interaction and evaluation, aesthetic judgments, and the jokes, images, and stereotypes pervading the mass media."[46]

By the time she wrote *Structure as the Subject of Justice*, Young seems to have "add[ed] some dimensions"[47] to this – in particular, Bourdieu's notion of "habitus," wherein bodily comportments, reactions, tastes, and preferences – stratified by class, wealth, and other socially salient categorizations – silently signal one's social position to others in such forms as voice, gesture, and a preference for, for example, scotch over beer (or vice versa). This represents an important complement to her account of -isms, showing how habitus, for example in the form of one's desire to find an apartment in a (white) middle-class neighbourhood ("where others like me live"), "(unconsciously) operates to reproduce structural inequalities" – where "structural inequalities" refer to "categorical inequalities, typically along the lines of class or class fraction, race, gender, ability, and sometimes ethnicity."[48]

The potential for social liberation from -isms – that is, for the elimination within contemporary society of racism, sexism, and so on – is available via Young's adoption of Giddens's conceptual tool of structuration. For much of

[40] K. Dowding, *Encyclopedia of Power*, op. cit., p. 255.

[41] I.M. Young, *Justice and the Politics of Difference*, op. cit., p. 133.

[42] Ibid., p. 146.

[43] Ibid., p. 147.

[44] Ibid., p. 151.

[45] Ibid., p. 146.

[46] Ibid., p. 148.

[47] See I.M. Young, *Structure as the Subject of Justice*, op. cit., p. 62.

[48] Ibid., p. 59.

the 20th century, social theorists had tended to coalesce around either an agent-centric paradigm, wherein individual actions are conceptualized as autonomous and largely unconstrained, or one that is structure-centric, in which structures of power constrain/determine human behaviour. Giddens's structuration, on the other hand, recognizes a duality: structure shapes human action, yet it is simultaneously and recursively constructed by those human actions. It is this latter aspect that suggests that humans have the capacity to alter their actions, to change their behaviours – including those actions and behaviours that reproduce -isms at the level of practical consciousness. Of course, the fact that they play out inadvertently poses a challenge: if humans are unaware of their racist, sexist, etc., tendencies, how can they correct them? The solution is to raise consciousness, to bring that which is inadvertent to the level of discursive consciousness – where discursive consciousness is understood as a level of experience where actors know what they are doing and can provide reasons for their behaviour. Consciousness-raising happens through social groups gathering to discuss their experiences of being treated as the Other – with recognition of common themes shared across their experience, and a vocabulary with which to describe it, emerging in their discussions. That occurred with the women's movement in the 1960s, and with the Black liberation movement in the late 1960s. Miranda Fricker provides an excellent example that sheds light on how consciousness-raising works: when women endured sexualized comments in the workplace etc., prior to the 1960s it was brushed aside as "flirting" or "harmless fun"; but when several women came together to discuss similar experiences, they began to develop a vocabulary around it – calling it "sexual harassment" – and eventually bringing/raising awareness of the wrongness of such behaviour to the level of men's discursive consciousness.[49]

Let us take stock. It should be apparent, at this point, that Young's critical theory grounds the justness or injustice of a social arrangement/outcome in a social theory of structural power. She provides a suite of concepts and tools that philosophers could draw upon to normatively reason about socially unjust outcomes – including, structural power; social groups and their experience of -isms; practical consciousness; consciousness-raising activities; and Giddens's structuration. Crucially, the incorporation of consciousness-raising spaces within her framework provides the mechanism for racist, sexist, etc., behaviours to be "named" – for example,

[49] M. Fricker, *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford University Press, Oxford 2007, pp. 150–151, https://doi.org/10.1093/acprof:oso/9780198237907.001.0001.

as “sexual harassment,” as we saw in Fricker’s account. Left unnamed, they go undetected – and the behaviours continue unabated, re-enacted in society (as was the case with the inappropriate, sexualized comments that were part of workplace culture before women’s groups called them out, and male managers were sent to workplace gender-awareness workshops). And, insofar as such behaviours are re-enacted in society, they are more likely to be reproduced in AI outcomes. It is worth highlighting the robustness of Young’s account. Not only does it give us the concepts with which to ideate racism, sexism, and other -isms – it also provides the tool that helps counter (and perhaps one day eliminate) them.

Furthermore, her consciousness-raising spaces can nourish the moral deliberations of philosophers. When morally relevant facts rise to discursive consciousness, philosophers have a broader array of facts to contemplate. Additionally, they gain the capacity to censure the behaviour of any perpetrators who, though now aware, are nonetheless unmoved.

5. AI Outcomes

I have shown that, when a social outcome implicates racism, sexism, and other -isms, an assessment of its injustices necessitates the use of a critical theory and an account of structural power. However, insofar as this approach tackles -isms *tout court*, advocacy for it would seem to hold even without AI.

Does anything change when we apply the approach to AI? Certainly, AI compounds the issue, reproducing those -isms in its outputs. Further, given the opacity of neural networks, we might not understand why that has happened (at least at the level of/inside the black box). However, the social theory within Young’s account allows us to better understand the social phenomena that caused the racist, sexist, etc., AI outputs. As we have seen, an integral part of Young’s critical theory is the value it gives to consciousness-raising spaces. Insofar as consciousness-raising helps curtail inadvertent sexism, racism, etc., and insofar as those -isms are moral wrongs that ought to be curtailed, it follows that consciousness-raising spaces ought to be developed and maintained to help identify and address instances of AI bias.

But how would consciousness-raising activities help here? How would they ameliorate the detection and redress of AI bias? Such bias sometimes only comes to light when historically marginalized people have a "hunch" that the algorithm is treating them differently. Without consciousness-raising activities, that hunch may remain undisclosed; it may even remain unidentified as a phenomenon – silently and unwittingly endured by marginalized people as "an inconvenience," rather than a form of "discrimination" or "harassment."[50] Consciousness-raising activities, on the other hand, provide a forum for discussing such hunches, sharing adverse experiences, and identifying patterns of AI bias. This process allows for the feedback of identified bias to AI developers. For example, African American and Hispanic communities could discuss the impacts of predictive policing and parole review AI systems on their lives; by sharing their individual experiences of (what at first may seem like) "unfortunate" parole denials, a pattern becomes discernible and (racial) bias becomes apparent.

One important question to consider is whether consciousness-raising activities replace existing mechanisms for addressing AI fairness, or do they complement them. Consider some existing mechanisms for addressing AI fairness:

- COMPAS, an AI tool used to predict recidivism, was shown to be biased against Black offenders – prompting the development of a race-neutral version of the algorithm.[51]
- Some companies deploy "gender decoders" to analyse job descriptions and detect subtle language biases that may deter women from applying – terms like "executes" or "competitive" might be flagged as masculine-coded.[52]
- To counteract the over-representation of certain groups in training data, re-sampling techniques may be used to ensure more balanced representation – as seen with facial recognition technologies.

Whilst these existing mechanisms may be effective to some extent, consciousness-raising activities can enhance their effectiveness by alerting AI developers to instances of AI bias and the need for such interventions.

[50] This is analogous to the experience for many women in the 1950s who faced inappropriate, sexualized behaviour from male colleagues. At the time, such behaviour was often dismissed as "flirting" and considered an "inconvenience" by some female colleagues. It was only later, through consciousness-raising activities and the sharing of experiences, that they came to recognize and identify these behaviours as "discrimination" and "sexual harassment."

[51] J. Angwin et al., *Machine Bias*, in: *Ethics of Data and Analytics: Concepts and Cases*, ed. K. Martin, Auerbach Publications, Boca Raton 2016, pp. 254–264.

[52] K. Crawford, T. Paglen, *Excavating AI: The Politics of Images in Machine Learning Training Sets*, "AI & Society" 2021, Vol. 36, pp. 1105–1116, https://doi.org/10.1007/s00146-021-01162-8.

That said, some existing AI fairness mechanisms face legal constraints because they often require access to sensitive attributes (such as gender or ethnicity) that privacy laws may ringfence.[53] In this context, consciousness-raising activities could offer a viable alternative. Instead of mining sensitive data to detect bias or demonstrate compliance with fairness standards, AI developers can engage with discursive consciousness-raising forums. These forums bring attention to biases related to gender, ethnicity, and other protected traits, allowing developers to identify issues through participant feedback[54] rather than through direct access to sensitive information.

Let's consider a potential objection to the analysis presented in this paper. The paper explored two possible approaches to appraising the justness or injustice of AI outcomes: Rawls's and Young's. A critic might ask: there are other abstract theories within political philosophy other than Rawls's – why consider his? Our answer is twofold. First, that we can't not consider him. His theory has become the dominant ideal theory in political philosophy over the past 50 years, shaping the thinking of many contemporary political philosophers. By engaging with Rawls, we interact with how a substantial portion in the field approach questions of justice and injustice. Second, AI scholars have already reached for Rawls's theory to answer questions about AI-exacerbated social injustice. Indeed, as Jørgensen and Søgaard note, "Researchers and industry developers in artificial intelligence (AI) and natural language processing (NLP) have uniformly adopted a Rawlsian definition of fairness."[55] One reason I assessed Rawls's theory within this paper was to make clear that it cannot answer the sorts of questions that worry many who study AI bias. Our analysis is intended to save them time and

[53] Yan et al. make this point too; see S. Yan, H.-T. Kao, E. Ferrara, *Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes*, "Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency" 2020, p. 1715, https://doi.org/10.1145/3340531.3411980.

[54] Of course, a form of feedback collection already exists within AI: a user may be presented with a short, in-app survey or with a prompt to rate the fairness of the app; an app may include a "Report bias" button; or the AI system might monitor user behaviour, noting that the user frequently overrides AI recommendations. But, whereas these existing mechanisms entail feedback from a single user, the consciousness-raising feedback is from many users, who, through the process of communicating their shared experience with one another, have clarified the phenomenon of group bias.

[55] A.K. Jørgensen, A. Søgaard, *Rawlsian AI Fairness Loopholes*, op. cit., p. 1185.

effort – steering them away from a philosophical path that cannot speak to the issues they seek to tackle within AI fairness.

Consider a second query about the paper. The critic might acknowledge that Young's theory indeed considers structural power, but then ask: but why should we? At one level, we can respond that, unless we do, we cannot grapple with those AI outcomes that implicate and reproduce structural power inequalities. But let's consider the critic's query more deeply. Perhaps they are saying that, insofar as structural racism, sexism, etc., are inadvertent, we cannot assign moral blame/ culpability to anyone for them – as such, we should ignore -isms in our deliberations about the moral permissibility of AI outcomes. My response is that this suggestion fails to grasp the interplay between Giddens's notion of structuration and the revelatory effects of consciousness-raising activities. The latter provides actors with information and insights that allow them to recognize their actions and reflect on them. The former shows us that agents retain agency – they *can* change their actions; and insofar as persons can change a morally impermissible or unjust action, we can hold them responsible – indeed, we could blame them, even (once we conduct appropriate moral deliberations that weigh any mitigating factors that could account for their inaction).[56]

6. Conclusion

Many scholars have engaged with Rawls's justice as fairness when studying AI fairness. We showed, though, that Rawls's theory, lacking a sociological theory of structural power, was not fit for that purpose – but that it was never intended for that purpose, either: it is supposed to move us towards Rawls's ideal version of justice, rather than to address, and move us away from, any particular -ism

[56] Tetyana Krupiyu argues that we ought to recognize the computer/data scientist's contribution to AI, rather than just thinking of the algorithm and its outputs, since this helps "capture the fact that computer scientists make subjective decisions in the course of creating the architecture that enables the AI decision-making process to collect, aggregate and analyse data. […] Often, the decisions of computer scientists are hidden and reflect a particular understanding of the world. For example, computer scientists make assumptions when deciding how to represent a person in a model" (T. Krupiyu, *A Vulnerability Analysis: Theorising the Impact of Artificial Intelligence Decision-Making Processes on Individuals, Society and Human Diversity from a Social Justice Perspective*, "Computer Law & Security Review" 2020, Vol. 38, 105429, https://doi.org/10.1016/ j.clsr.2020.105429, p. 8 of 25).

injustices. This revelation allowed us to conclude that AI ethicists should not look to Rawls when they ask questions about AI decisions that are racist, sexist, etc.

On the other hand, we showed that Young's approach, drawing on a sociological theory of structural power, is well-suited to the task. Her concept of practical consciousness, as we saw, accounted for unspoken, pernicious aspects of racism, sexism, and so on. Moreover, Young's device of consciousness-raising activities, as I showed, can illuminate and "name" unjust behaviours. That can, as I argued, nourish philosophers' moral reasoning about AI outcomes that are racist, sexist, etc. It can, also, as we saw, help remove the racism, sexism, and other -isms that get reproduced in AI outcomes.

Bibliography

Anderson E., *What Is the Point of Equality?*, "Ethics" 1999, Vol. 109, No. 2, pp. 287–337.

Angwin J., Larson J., Mattu S., Kirchner L., *Machine Bias*, in: *Ethics of Data and Analytics: Concepts and Cases*, ed. K. Martin, Auerbach Publications, Boca Raton 2016, pp. 254–264.

Barsotti F., Koçer R.G., *MinMax Fairness: From Rawlsian Theory of Justice to Solution for Algorithmic Bias*, "AI & Society" 2024, Vol. 39, pp. 961–974, https://doi.org/10.1007/s00146-022-01577-x.

Bay M., *Participation, Prediction, and Publicity: Avoiding the Pitfalls of Applying Rawlsian Ethics to AI*, "AI and Ethics" 2024, Vol. 4, pp. 1545–1554, https://doi.org/10.1007/s43681-023-00341-1.

Binns R., *Algorithmic Accountability and Public Reason*, "Philosophy and Technology" 2018, Vol. 31, pp. 543–556.

Buolamwini J., Gebru T., *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, "Proceedings of Machine Learning Research" 2018, Vol. 81, pp. 1–15.

Buranyi S., *Rise of the Racist Robots: How AI Is Learning All Our Worst Impulses*, "The Guardian," 8.08.2017, URL: https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses.

CBC Radio, *Police Are Considering the Ethics of AI, Too*, 21.09.2018, URL: https://www.cbc.ca/radio/spark/tech-in-policing-1.4833189/police-are-considering-the-ethics-of-ai-too-1.4833194.

Crawford K., Paglen T., *Excavating AI: The Politics of Images in Machine Learning Training Sets*, "AI & Society" 2021, Vol. 36, pp. 1105–1116, https://doi.org/10.1007/s00146-021-01162-8.

Daniels N., ed., *Reading Rawls: Critical Studies on Rawls' A Theory of Justice*, Stanford University Press, Stanford 1975.

Dastin J., *Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women*, Reuters, 9.10.2018, URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Dowding K., *Encyclopedia of Power*, SAGE, Thousand Oaks 2011.

Dunn E., *Public Attitudes to Data and AI: Tracker Survey*, Centre for Data Ethics and Innovation, London 2022.

Farrelly C., *Justice in Ideal Theory: A Refutation*, "Political Studies" 2007, Vol. 55, pp. 844–864.

Ferrara E., *Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies*, "Sci" 2024, Vol. 6, No. 1, pp. 1–15, https://doi.org/10.3390/sci6010003.

Fricker M., *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford University Press, Oxford 2007, https://doi.org/10.1093/acprof:oso/9780198237907.001.0001.

Gabriel I., *Toward a Theory of Justice for Artificial Intelligence*, "Daedalus" 2022, Vol. 151, No. 2, pp. 218–231, https://doi.org/10.1162/daed_a_01911.

Giddens A., *Central Problems in Social Theory: Action, Structure, and Contradiction in Social Analysis*, MacMillan Education, London 1979.

Giddens A., *The Constitution of Society: Outline of the Theory of Structuration*, Polity, Cambridge 1986.

Heidari H., Ferrari C., Gummadi K., Krause A., *Fairness behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making*, "Advances in Neural Information Processing Systems" 2018, Vol. 31.

Jørgensen A.K., Søgaard A., *Rawlsian AI Fairness Loopholes*, "AI and Ethics" 2022, Vol. 3, pp. 1185–1192, https://doi.org/10.1007/s43681-022-00226-9.

Krupiy T., *A Vulnerability Analysis: Theorising the Impact of Artificial Intelligence Decision-Making Processes on Individuals, Society and Human Diversity from a Social Justice Perspective*, “Computer Law & Security Review” 2020, Vol. 38, 105429, https://doi.org/10.1016/j.clsr.2020.105429.

MacMillan D., *Eyes on the Poor: Cameras, Facial Recognition Watch over Public Housing*, “The Washington Post,” 16.05.2023, URL: https://www.washingtonpost.com/business/2023/05/16/surveillance-cameras-public-housing/.

McNay L., *Recognition as Fact and Norm: The Method of Critique*, in: *Political Theory: Methods and Approaches*, eds. D. Leopold, M. Stears, Oxford University Press, Oxford 2008, pp. 85–105.

Mills C.W., *“Ideal Theory” as Ideology*, “Hypatia” 2005, Vol. 20, No. 3, pp. 165–184.

Mills C.W., *Retrieving Rawls for Racial Justice? A Critique of Tommie Shelby*, “Critical Philosophy of Race” 2013, Vol. 1, No. 1, pp. 1–27.

PBS Frontline, *A Class Divided*, “CosmoLearning” 1985.

Rafanelli L.M., *Justice, Injustice, and Artificial Intelligence: Lessons from Political Theory and Philosophy*, “Big Data and Society” 2022, Vol. 9, No. 1, https://doi.org/10.1177/20539517221080676.

Rawls J., *The Law of Peoples*, Harvard University Press, Cambridge, MA, 1999.

Rawls J., *A Theory of Justice*, Harvard University Press, Cambridge, MA, 1971.

Reuters, *Amazon Ditched AI Recruiting Tool that Favored Men for Technical Jobs*, “The Guardian,” 11.10.2018, URL: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine.

Simmons A.J., *Ideal and Nonideal Theory*, “Philosophy & Public Affairs” 2010, Vol. 38, pp. 5–36.

Weidinger L., McKee K., Everett R., Huang S., Zhu T., Chadwick M., Summerfield C., Gabriel I., *Using the Veil of Ignorance to Align AI Systems with Principles of Justice*, “Proceedings of the National Academy of Sciences of the United States of America” 2023, Vol. 120, e2213709120, https://doi.org/10.1073/pnas.2213709120.

Yan S., Kao H.-T., Ferrara E., *Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes*, “Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency” 2020, pp. 1715–1724, https://doi.org/10.1145/3340531.3411980.

Young I.M., *Communication and the Other: Beyond Deliberative Democracy*, in: *Democracy and Difference: Contesting the Boundaries of the Political*, ed.

S. Benhabib, Princeton University Press, Princeton 1996, pp. 120–136, https://doi.org/10.1515/9780691234168-007.

Young I.M., *Equality of Whom? Social Groups and Judgments of Injustice*, “The Journal of Political Philosophy” 2001, Vol. 9, No. 1, pp. 1–18.

Young I.M., *Justice and the Politics of Difference*, Princeton University Press, Princeton 1990.

Young I.M., *Structure as the Subject of Justice*, in: I.M. Young, *Responsibility for Justice*, Oxford University Press, Oxford 2011, https://doi.org/10.1093/acprof:oso/9780195392388.003.0002.

Zhang M., *Google Photos Tags Two African-Americans as Gorillas through Facial Recognition Software*, “Forbes,” 1.07.2015, URL: https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/.