Why Moral Agreement is Not Enough to Address Algorithmic Structural Bias

Paige Benton[1]

University of Pretoria, South Africa

Abstract. One of the predominant debates in AI Ethics is the worry and necessity to create fair, transparent and accountable algorithms that do not perpetuate current social inequities. I offer a critical analysis of Reuben Binns's argument in which he suggests using public reason to address the potential bias of the outcomes of machine learning algorithms. In contrast to him, I argue that ultimately what is needed is not public reason per se, but an audit of the implicit moral assumptions of societies within which algorithms are built and applied.
Public justification is appealing since it offers us the possibility to align the decision-making outcomes of the algorithm with the core moral values of stakeholders within a constitutional democratic society. My concern is that the common moral principles that form the foundation of public reason are not necessarily neutral, as they still express specific moral ideals and normative standards even if there is moral agreement by society as a whole, or among different stakeholders within society.
Appealing to such normative standards may thus still lead to algorithmic outcomes being biased as common moral values may very well still be discriminatory even though they are formed from a consensus, and even if public reason is applied as a kind of filter for potential algorithmic outcomes. Hence, I argue that these implicit moral norms within society that we take as a given in public reasoning, need to be audited from generation to generation in order to effectively mitigate potential algorithmic bias.

Keywords: Moral Agreement · Public Reason · Ethics of AI · Algorithmic Bias ·Accountability.

1 Introduction

Due to the rise of machine learning and the complex and intimate decisions generated by machine models there is a growing need for reflecting on the ethical implications and ethical parameters of the outcomes of machine learning algorithms. These concerns have gained increased awareness in AI in the recent years with the rise of groups such as FAT/ML [8], DADM [26], and the Algorithmic Justice League [34] to name a few. Organizations such as these shed necessary light on issues such as the transparency of decision-making, algorithmic bias, identifying discriminatory data policies and prejudicial training datasets.

One of the core ethical concerns in machine learning, the focus of this paper, is the moral prescription embedded in machine learning algorithms. Whose morality matters and what moral values are important? Should the morality of the computer programmer count, or the moral values of the company creating the algorithm, or the values of the country in which the algorithm is created? Questions such as these illuminate the central concern, namely, which moral values are being endorsed and does this promotion lead to discrimination and bias?

Reuben Binns [2], suggests turning to political philosophy for a solution. Binns proposes applying the method of public justification to analyse the ethical risks and quality of potential outcomes of the algorithms at issue. Turning to public reason is appealing since it offers us the possibility to align the decision-making outcomes of the algorithm with the core moral values of stakeholders within a liberal constitutional democratic society, thus legitimising the decision-making outcomes of the algorithm itself.

My concern is that the common moral principles that form the foundation of public reason in themselves are not necessarily neutral. On the contrary, they express specific moral ideals and normative standards – i.e., liberal democratic moral norms – even if moral agreement (by society as a whole, or among different stakeholders within society) seems an attractive ideal as a foundation for the justification of algorithmic outcomes in the context of AI Ethics. Appealing to such normative standards may still lead to excluding some at the expense of others and encouraging some virtues while neglecting others. There thus still exists the possibility for algorithmic outcomes to be biased as common moral values may very well still be discriminating even though they are formed from a consensus, and even if public reason is applied as a kind of filter for potential algorithmic outcomes.

Hence, I argue that societal consensus is not enough to ensure the method of public reason as an adequate purifier of algorithmic outcomes as it is not adequate for legitimising moral values. Given that it is these implicit moral norms within society that we take as a given in public reasoning, I argue they need to be audited from generation to generation in order to expose and critically evaluate implicit moral societal assumptions that may be deemed harmful only in a hindsight evaluation of a liberal constitutional democratic society. Note, this paper does not offer technical solutions to the problem of algorithmic bias but is a critical analysis of Binns's article from the lens of political philosophy and the ethics of AI.

In order to illustrate the need for auditing the implicit moral assumptions of public reason, namely liberal democratic moral norms, Section 2 is a brief account of structural bias in machine learning algorithms. Section 3 is an overview of the justificatory process of public reason that enables societal consensus in moral pluralistic societies. Thereafter Section 4 examines the potential risks for the reproduction of moral bias by machine learning algorithms using the method of public justification. Lastly, Section 5 offers a concluding evaluation of moral agreement as a basis for justification.

2 Structural Bias and Critical Machine Learning

There is a growing need to re-evaluate the ethical implications that result from algorithmic decision-making in machine learning. Firstly, a significant concern is the problem of pre-existing or historical bias[1] being perpetuated via the outcomes of machine learning algorithms. Bias in this sense is structural bias and refers to the development of prejudiced judgements based on preconceived views that are informed by societies current cultural assumptions and systemic injustices [11,12]. Pre-existing bias[2] in the context of AI is a problem when decision-making algorithms mimic and propagate the social injustices that are already evident in societies [21]. The reasons for this propagation are diverse.

Essentially, in developing algorithms developers construct boundaries, rules and success definitions for their algorithm, all of which act as both moral and practical constraints. Constraints are ‘practical’, in the sense that algorithms need parameters and rules in order to delineate relevant information from irrelevant information. They are ‘moral’, in that they seemingly embody normative prescriptions of the developer, insomuch as they delineate goals, embed values and ideological assumptions when modelling the algorithm, and also as they reflect and potentially amplify existing societal bias in training data [4,1]. The practical and moral justification for the exclusion or inclusion of information creates what O’Neil terms ‘moral blind spots’, implying that developers construct algorithms in line with their preconceived societal judgements [24]. This is precisely why O’Neil claims “models are opinions embedded in mathematics” [24].

One of the dangers in the context of machine learning is that the justification underlying the decision-making algorithm is implicit and as such their moral reasoning may appear as “black boxes” [12, 24, 25]. As seen in [13] there is movement within the industry to make the constraints and success definitions of algorithms available to the public. Thinkers such as Crawford [6], O’Neil [24], and Nissenbaum [23], argue that it is not enough to publicise the underlying justification, but instead suggest, just as the training set, the source of the data the model learns from, needs to be publicly audited, to make sure that the data itself is unbiased, so too should the success definition of the algorithms be audited, to shed light on the moral assumptions underlying algorithms [24]. This opens the possibilities for widening the requirements of a successful algorithm to include fairness as a metric as opposed to the current metrics of efficiency and profit [24].

[1] Note critical machine learning focuses on a board range of issues, such as fairness, transparency and accountability in machine learning processes. The focus of this paper is only on bias.

[2] Note that there are a wide variety of different kinds of biases in AI ethics. For example, technical, emergent and representational bias to name a few. The focus of this paper is preexisting bias since this form of bias addresses systemic societal injustices which can be propagated by implicit societal moral norms. For future detailed discussion of the forms of bias see
[21] pp 4-7.

Even instituting fairness as a metric to measure success of algorithms is difficult as after all, on the one hand there is no standard interpretation of fairness, while on the other hand the interpretation would be dependent on the various contexts in which models are used [3,21,22]. Given the plurality of interpretations and understanding of moral and political norms, this calls for consensus on norms vis-à-vis., accountability, privacy and fairness [22]. Without societal consensus ethical guidelines and frameworks being suggested in AI ethics are ineffective. [15] echoes this concern when arguing that ethical guidelines offer superficial aid to ethical concerns in the AI industry, considering that current guidelines are voluntary and not an obligation, nor are there consequences for those that do not uphold the principles. Similarly, ethical frameworks in AI are so vast that they create a rise of 'ethics shopping' insofar as developers can choose frameworks that align with their own agenda, making the guidelines ineffective [9]. As a result, ethical norms and parameters in AI are currently bound to relativistic debate and implementation.

Binns [2] offers an attractive solution to try reach moral agreement in AI ethics. Using the method of public justification, he claims that consensus can be formed regarding core moral values, which we turn to now.

3 Public Reason

Prior to discussing Binns's argument for the use of public reason in algorithm accountability, allow me some background remarks on public reason and its use in political philosophy and liberal theory.

In essence, public reason as a method of justification for moral and political principles requires that these principles are publicly justifiable and accepted by those to whom these principles apply [27]. This tradition of moral agreement of overarching principles that can act as a moral consensus for societal rules traces back to thinkers such as Hobbes [17], Kant [19], Rosseau [33] and more recently to Rawls [30] and Habermas [16]. Public reason is intrinsically tied to liberal theory, since a key premise of this method of justification is the conception of persons as free to construct their own social, moral and political norms. Considering that persons have the freedom to construct their own moral views, this gives rise to the phenomenon of moral pluralism in constitutional liberal democratic societies. Moral pluralism is a notion used by liberal theorists such as Rawls to denote an essential feature of liberal societies i.e., moral disagreement regarding fundamental moral principles, ideals and values. The 'burdens of judgement' in a liberal society signify that we need to allow room for reasonable disagreement among moral values yet develop moral consensus on political principles [30].

Consequently, moral truths (i.e., universal moral views namely religious, metaphysical or philosophical doctrines that prescribe comprehensive moral values) cannot be the basis of moral agreement in liberal societies since all citizens hold contrasting and conflicting views on what the foundation of moral truth should be [30]. The only type of moral agreement citizens can hope for is to agree to political principles and values that all citizens could indorse and would consider morally reasonable for all citizens to follow given the circumstance of moral pluralism. To put it another way, citizens in acknowledging their diverse moral values and ideals reach agreement on political principles and values that align with and promote their own values.

This point of ‘moral congruence’, Rawls refers to as the ‘overlapping consensus’ [30]. An overlapping consensus is achieved when the same conception of justice receives public support from diverse comprehensive doctrines. Keep in mind though, as stated above, that all citizens affirm the same political principles, yet do so for different reasons, influenced by the different moral doctrines that they adhere to [29]. These political principles that citizens agree on then limit the boundaries of what is acceptable moral disagreement within a liberal constitutional democracy.

Public reason requires that political principles can be affirmed by persons upholding multiple moral doctrines for the principles do not presuppose an antecedent truth of any one moral doctrine. But if moral truth is not the foundation for political principles, and citizens of liberal society hold conflicting moral values, what is the underlying foundation of the political principles that they all can endorse?

Rawls’s seminal account of public reason in ‘Justice as Fairness’ introduces the method of ‘reflective equilibrium’[3] to justify the underlying common moral norms that could be the foundation for political principles of a liberal society. Broadly speaking, reflective equilibrium is a method to expose and align the moral assumptions (i.e., ‘moral sentiments’) and acceptable political principles (i.e., ‘considered judgements of justice’) of citizens of a constitutional democracy [31,32]. A state of wide equilibrium is reached when, after continual reflection, implicit moral values that are common to all citizens are identified.

One of the core implicit moral values is freedom of conscience. All citizens would recognize themselves and others as having the freedom to choose how to give their life meaning, given the premise that liberal societies are characterised by moral disagreement. Other implicit moral norms in a liberal society include toleration, equality, and freedom to name a few. These are all core implicit moral norms as they help facilitate citizens individual freedom.

These implicit moral values act as points of convergence of all citizens’ conflicting moral doctrines, and if citizens can agree to uphold these implicit moral norms then these norms become the underlying constitutional essentials for the political principles [31]. Given this, citizens can develop moral agreement without appealing to moral doctrines directly, but rather by appealing to these implicit moral standards that are collectively shared across all moral doctrines to which they subscribe. Political principles that best manifest these common implicit moral values are principles that all persons would agree to. Principles such as freedom of thought, freedom of movement, freedom of occupation, and equal opportunity to wealth are examples of the kinds of political values that citizens would find reasonable to uphold [28].

Allow me a hypothetical example to illustrate my point. A citizen belonging to the Islamic or Christian religion may choose to uphold the same political virtue of toleration or choose to uphold the value of liberty of conscience, as a citizen who subscribes to Atheism, as these values and principles themselves help to secure both

[3] Although Rawls popularized the term ‘reflective equilibrium’ and applied this method to the field of political philosophy, this method of justification was development prior to him by Nelson Goodman – see *Fact, Fiction, and Forecast* [14].

citizens' ability to practice their moral doctrine. There can be variants of both Christianity, Islam, and any other religious, metaphysical or philosophical moral doctrines. However, the radicalized versions of any moral doctrine will not be acceptable in a liberal society since they would not align with the implicit moral norms of a liberal society such as liberty of conscience. The significance is that only liberal versions of any moral doctrine can be aligned with and encouraged by the implicit moral norms themselves. This moral prescription is necessary for the possibility of moral agreement in a liberal society, the implications of which are discussed in the following section. For now, it is important to turn to Binns's argument.

Binns [2] in acknowledging the plurality of moral views between various stakeholders, recognizes the problem that some stakeholders may not be accepting of the moral justification used by the developers when constructing any algorithm. As such, this would lead to a further rise in moral disagreements and widespread debate as to what method and values should be used when modelling algorithms. To circumvent this disagreement, Binns, as previously stated, turns to public reason to develop moral consensus between those that model algorithms (i.e., 'decision-maker') and those to which algorithms apply (i.e., 'the decision-subject'). Since the 'decision-maker' and the 'decision-subject' exist in a society characterised by freedom of conscience, the possibility for moral consensus seems unlikely, since, by definition, moral pluralism makes moral agreement appear less plausible. This situation could lead to a standoff between prioritising either the 'decision-maker's' or the 'decision-subject's' moral values and interests. Either way this prioritisation would lead to some disenfranchised stakeholders [2].

Binns suggests using public reason as the gate keeper of accountability when modelling algorithms. For Binns, public reason as a method of justification can be applied between 'decision-makers' and 'decision-subjects' to mediate reasonable ethical and epistemic standards from the unreasonable, by identifying the overarching common values all persons can agree to. This agreement then sets the boundaries for acceptable rules in AI.

The advantages of public reason in AI, according to Binns, are the following: Firstly, it can help identify problems of algorithmic bias by identifying biases in the data or in the modelling, whilst comparing the moral values represented in bias incidents to determine if they are unreasonable by reference to the implicit moral norms underpinning liberal societies. Secondly, if stakeholders agree to collective moral norms they find reasonable to guide the industry, this could legitimise the decision-making aspect of modelling algorithms. Thirdly, public reason may be a useful method of justification to denote public and private algorithmic accountability, since different values, scope and context to which an algorithm applies may be the reason for an action being discriminatory or reasonable. Subsequently, Binns suggests public reason could be used to determine reasonable epistemic standards from the unreasonable, where there is debate over correlation, causation and the nature of a relation between entities. Lastly, public reason could be used to delineate common moral values from moral doctrines, while making sure that the latter are not used as a justification for preferential treatment in the modelling of algorithms or as grounds for objection to algorithms [2].

Binns acknowledges two problems: Firstly, the method of public reason is currently being used in AI seeing as general ethical frameworks and recommendation policies such as AI4People [10] align their guidelines with legal and political principles that govern liberal societies. Secondly, as mentioned in the previous section, opacity of the decision-making aspect of algorithms implies that it may prove impossible to know if the justification of algorithms aligns with the moral values in a liberal society. Yet, Binns claims the goal of algorithms and their training data can still be assessed in light of public reason [2].

Binns's argument for public reason to guide accountability in AI is attractive as it provides a solution to the continual disagreement between stakeholder's interests and values. However, a shortcoming of public reason is that it is only applicable to liberal societies, as previously mentioned. The implication of this is that it cannot act as the method for public justification on a global scale, since not all countries subscribe to liberalism. Considering that there are countries that are modelling and training algorithms with their own moral doctrines guiding the decision-making process, it seems that a universal method of justification for AI accountability, founded on public reason is less likely than Binns perhaps suggests. Without all countries agreeing on the same implicit moral norms, there is no basis for moral agreement in public reason. However, on a national and supranational level public reason remains a possibility as long as it is among liberal societies with the shared moral values see 'Artificial Intelligence: The Global Landscape of Ethics Guidelines' [18].

Acknowledging the problem that public reason is only applicable to liberal societies — since it requires persons to uphold liberty of conscience as an implicit moral norm — illustrates that there is moral prescription encapsulated into the content of public reason. Hence, liberal societies are not neutral or accommodating of all moral values. Instead, the values that are accommodated are those that are liberal in nature. I will address why this is a concern in more detail in the following section.

4 The Necessity to Audit Implicit Moral Norms

As established in the previous section, toleration, equality and freedom of conscience are some of the core implicit moral values of a liberal society. Moral values such as these promote political principles such as the absolute[4] respect of individual rights, freedom of speech, equal opportunities and tolerance to those citizens which hold alternative moral doctrines. Due to this, the rights and liberties of persons must be given a prioritised position of importance in that they have to be instituted and affirmed unconditionally. Hence, the political principles they promote are not prejudice against persons based on their race, gender, sexual orientation or religious affiliation. In fact, these moral values and political principles encourage inclusion of dissimilar persons and their freedom to develop, pursue and follow their own ends in line with their chosen moral doctrine.

[4] Rights remain absolute insofar as they are the prioritised values in liberal constitutional democracies.

If these implicit moral norms are not inherently prejudicial why did I suggest earlier that it would be necessary calling for a continual audit of these norms of a liberal society? The reason is precisely because these moral values are just that, prescriptive notions of good and bad ideals that have become normalised, and used as the moral foundation against which we distinguish morally permissible from morally impermissible actions.

Liberal societies will necessarily include some values at the expense of others; that is the nature of morality itself. Unavoidably theorists, when developing moral or political theories, cannot account for all possible forms of life, there are certain moral assumptions that have to be made, and these assumptions will then delineate the permissible from the impermissible moral actions. Thus, any political system will naturally favour, and hence encourage certain societal ideals, and as a result would exclude other moral ideals.

Thus, liberal values are, once again, not neutral values. Instead, the implicit moral norms of liberal society become societal moral 'blind-spots', insofar as citizens are encouraged by their social context, and reenforce the values that surround them. Similarly, stakeholders (those modelling algorithms and those impacted by algorithms) in AI industry make decisions from within their social, moral and political context. The societal moral norms of our time will inform what we think is the correct moral judgements to make. Even if public reason is used to clarify reasonable moral values for potential algorithmic outcomes, the implicit moral values that underly this societal consensus still promote moral ideals. In order to be cognizant of the moral ideals promoted in liberal constitutional societies it is important to audit the values and their implementation to assess the ostracizing impact of these values, which may be deemed harmful only in hindsight evaluation.

To illustrate this point, let us return to the hypothetical example from the previous section. To recap, a feature of liberal society is moral pluralism. Insomuch as persons wish to affirm their own freedom to choose their own moral doctrines and life plans, the foundation for moral agreement (i.e., public reason) is freedom of conscience. In recognising freedom of conscience as the implicit moral norm of liberal societies, persons indirectly accept that their moral doctrine is not the only doctrine that can exist in a liberal society. Hence, by agreeing to this, citizens acknowledge and reaffirm the moral view that there are multiple moral truths. To put it another way, citizens of a liberal society have to concede that there cannot be only one moral truth i.e., one moral doctrine that should govern the social and political context. Thus, even if citizen A subscribes to Islam, citizen B subscribes to Judaism and citizen C subscribes to Atheism, all citizens have to be conscious that their fellow citizens have the right to choose their own moral doctrine. Given this, citizen A, B and C cannot hold radicalised or fundamentalist versions of their moral doctrines, owing to the fact that by definition a radicalised version of any philosophical, metaphysical or religious doctrine implies that there is only one moral truth, and that all persons should uphold this antecedent notion of moral truth. An example of political system founded on this form of reasoning is a theocracy, which promote a religious moral truth. Therefore, citizens of liberal democracies that would subscribe to religious fundamentalism would not identify with the implicit moral norms of liberalism (such as liberty of conscience), hence they are not able to reach moral agreement via the method of public reason. Considering that such citizens could not publicly justify their moral doctrine, their moral doctrines would be considered unreasonable in a liberal society.

This suggests that the kinds of moral doctrines that can flourish in a liberal society are purely *liberal* moral doctrines. That is to say liberal citizens can have any religious affiliation if and only if they denounce the radicalised version of their moral doctrine. The importance of this is that although liberal society is not prejudiced against religious affiliation, there are moral doctrines that cannot gain support via public reason.

What happens to individuals in liberal societies who hold radicalised moral doctrines? I argue that there is a potential tendency for algorithmic outcomes to exclude or censor such individuals, on social media platforms such as Twitter, Instagram and Facebook, on two grounds, firstly for instances of hate speech and secondly for not upholding liberal values.

In the infancy of social media platforms, public discourse was not hampered by regulations regarding when a post represents freedom of speech or when it transitions to hate speech. With the rise of increasing incidents of xenophobia, racism, sexism, bigotry and harassment, platforms have altered their community guidelines to try address these intolerances and others. In addressing these intolerances many social media platforms are starting to rely on algorithms to detect and censor user's content before it becomes available to other users and causes harm [5]. Although these algorithms are in their early stages their potential to identify these discriminatory acts is alluring. Yet currently certain forms of discrimination are harder for algorithms to identify then others. For example, Facebook's algorithm identifies and removes only 38% of hate speech incidents as opposed to 99.5% of terrorist activity and 86% of violent images [20].

What these figures illustrate is the difficulty in identifying hate speech from freedom of speech. Take for example a social media user in a liberal constitutional society tweets: The only God that exists is Allah all other religious Gods are false Gods. Does this moral view express a liberal moral doctrine or a radicalized moral doctrine? One could argue that it appears to be an expression of this user's moral point of view. Alternatively, the fact that the user acknowledges and makes the statement that no other Gods exist, is implying an antecedent moral truth claim. If it were the latter, it would be problematic as the statement then contradicts the liberal value of liberty of conscience.

The question we need to ask is: Does censoring the above kinds of statements – that may not seamlessly align with liberal values – lead to protecting individual liberty or the creation of what would become historical bias and discrimination in hindsight evaluation? I suggest that only by critically auditing the implicit liberal moral values and their interpretation on an institutional and societal level would policy makers, ethicists, computer scientists, private and public corporations be able to arrive at an informed understanding of the kinds of biases and discrimination that could be promoted as a result of encouraging liberal values.

In view of the above, even if public reason is an appealing method of justification to gain moral agreement between stakeholders in the AI industry for liberal societies on a national and supernational level, the fact is that public reason, as a filter for identifying implicit moral values to determine reasonable algorithmic outcomes, needs to be reenforced by the continual auditing of the implicit moral norms that constitute the content of public reason. It is only through critical and continual examination of implicit moral values that we can hope to become aware of the implicit moral assumptions of liberal society. Current societal values are the hardest to identify as they are normalised, yet the most important to audit or expose, precisely because we have normalised them; after all, they naturally become the 'moral blind-spots' of liberal democratic societies.

5 Conclusion

How to combat historical prejudices and its proliferation when modelling algorithms is a huge hurdle in AI. This article aims to show that some of these hurdles can be overcome with the use of public reason as suggested by Binns. Public reason offers the attractive possibility of gaining societal consensus in constitutional liberal democracies characterised by moral pluralism. I argued however that the dilemma remains that public reason can only be the foundation of moral agreement in liberal societies, since not all countries subscribe to liberalism and uphold the moral value of liberty of conscience. Hence, many countries do not have the shared moral norms necessary for the foundation of agreement. Due to this, currently public reason remains a viable method of justification for regional, national and supranational agreement only.

This, points to the question: Is a global ethics framework centered on public reason possible, considering the moral disparity between countries? No matter the type of societal agreement, the concern remains that although persons may have shared moral norms and may agree to uphold moral values this agreement alone is not sufficient to legitimize the moral values themselves. Moral norms can still be discriminatory even if shared by all persons as I have shown above.

Accordingly, even if all algorithms are modelled in line with liberal constitutional democratic moral norms, these norms are in fact not neutral since they are still morally prescriptive. As such they prescribe liberal moral doctrines and values. This is precisely why one should never ask the question: is it possible to develop neutral and unbiased algorithms? Since no algorithm can be morally neutral. There is always a moral assumption embedded into its framework. Therefore, one should rather ask: how do we model algorithms in line with the implicit moral norms of said society, and does the institutionalization of these norms promote bias or discriminate either directly or indirectly? Given this, I argue, it is important to re-enforce the notion that moral agreement via public reason does not imply moral neutrality which is currently not acknowledged in the Ethics of AI.

The impossibility of moral neutrality, shows then that the potential for the perpetuation of historical bias does not end with the implementation of public reason. Instead, I suggest that only by AI stakeholders constantly reflecting and evaluating the societal moral norms and the moral assumptions underlying public reason, would it be possible for bias and discrimination to decrease. After all, the greatest obstacle to algorithmic accountability is the normalization of current societal moral values. Auditing the moral values of liberal democratic societies not only helps to identify implicit moral 'blind spots', but also helps to identify when moral opinions stated in public social media forums are reasonable, thereby reducing the potential of unreasonable censorship when modelling algorithms. A further point of research would be unpacking how the auditing of liberal societal moral norms should be undertaken and its implications for both public and private sectors of societies. Although the method of auditing moral norms requires further explanation (that is not possible here), it is apparent that such actions need the expertise of social scientists, especially philosophers, this paper is thus in the final instance, also a definite plea for the recognition by the tech community that inter-and multi-disciplinary collaboration is of core importance in the domain of the ethics of AI.

References

1. Barocas, S., Selbst, A.: Big data's disparate impact. California Law Review 104, 671–732 (2016)
2. Binns, R.: Algorithmic Accountability and Public Reason. Philos Technol 31, 543-556 (2017). https://doi.org/10.1007/s13347-017-0263-5 last accessed 2021/09/20
3. Binns, R.: Fairness in Machine Learning: Lessons from Political Philosophy. In Proceedings of Machine Learning Research. Friedler, S., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 81, pp. 1-11. Journal of Machine Learning Research (2018).
4. Brundage, M.: Limitations and Risks of Machine Ethics. Journal of Experimental and Theoretical Artificial Intelligence 26, 3, 355-372 (2014). http://dx.doi.org/10.1080/0952813x.2014.895108, last accessed 2021/09/09
5. Cobbe, J.: Algorithmic Censorship by Social Platforms: Power and Resistance. Philos. Technol. (2020). https://doi.org/10.1007/s13347-020-00429-0
6. Crawford, K.: The Trouble with Bias. NIPS 2017 Keynote. https://www.youtube.com/watch?v=fMym_BKWQzk ccessed 2021/09/20
7. Fairness, Accountability and Transparency in Machine Learning (2014) https://www.fatml.org Accessed 2021/09/20.
8. Floridi, L.: Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. Philos. Technol. 32, 185-193 (2019). https://doi.org/10.1007/s13347-019-00354-x, last accessed 2021/07/19.
9. Floridi, L., Cowls, J., Beltrametti, M. et al.: AI4 People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines 28, 689-707 (2018). https://doi.org/10.1007/s11023-018-9482-5, last accessed 2021/09/08.
10. Friedman, B., Brok, E., Roth, K.S., et.al.: Minimizing Bias in Computer Systems 28,1, 48-51. (1996). https://doi.org/10.1145/249170.249184, last accessed on 2021/09/20
11. Gianfagna, L., Piccarozzi, D. , Di Cecco, A.: "Explainable AI": Who Takes the Decisions For Us? (2019). https://towardsdatascience.com/explainable-ai-who-takes-the-decisions-for-us-97b1d33edd91, last accessed 2021/09/20.

12. Goodman, B., Flaxman, S.: European Union Regulations on Algorithmic Decision-Making and a Right to Explanation (2016). arXiv [stat.ML]. http://arxiv.org/abs/1606.08813, last accessed 2021/09/20.
13. Goodman, N.: Fact, Fiction, and Forecast. Cambridge, Massachusetts: Harvard University Press (1955).
14. Hagendorff, T.: The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 99-120 (2020). https://doi.org/10.1007/s11023-020-09517-8, last accessed 2021/09/20
15. Habermas, J.: Reconciliation Through the Public use of Reason: Remarks on John Rawls's Political Liberalism, The Journal of Philosophy, 92,3, 109–131 (1995).
16. Hobbes, T.: Leviathan. Rev. student edn, edited by R. Tuck. Cambridge, Cambridge University Press (1996).
17. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2, last accessed 2021/09/09
18. Kant, I.: Groundwork for the Metaphysics of Morals, translated by A. Wood. New Haven and London, Yale University Press (2002).
19. Koebler, J., Cox, J.: The impossible job: inside Facebook's struggle to moderate two billion people. Vice. 2018. https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works, last accessed 2021/09/10.
20. Mehrabi, N., Morstatter, F., Saxena, N., et.al.: A Survey on Bias and Fairness in Machine Learning.ACM Computing Surveys (CSUR) 54, 1-35 (2021).
21. Mittelstadt, B. Principles Alone Cannot Guarantee Ethical AI. Nat Mach Intell 1, 501–507 (2019). https://doi.org/10.1038/s42256-019-0114-4, last accessed 2021/09/05.
22. Nissenbaum, H.: How computer systems embody values. Computer, 34(3), 120–119. (2001).
23. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. 1st edn. CROWN, New York (2016).
24. Pasquale, F.A.: Restoring Transparency to Automated Authority. Journal on Telecommunications and High Technology Law, 9, Seton Hall Research Paper No. 2010-28, pp 235-256 (2011). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1762766, last accessed 2021/09/20
25. Pedreschi, D., Salvatore, R., Turini, F.: Discrimination-aware data mining. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 560-568. (2008). http://doi.org/10.1145/1401890.1401959, Accessed 2021/09/20.
26. Quong, J.: Public Reason. (2013). https://standford.library.sydney.edu.au/entries/public-reason/ , last accessed 2021/09/03
27. Rawls, J.: The Basic Liberties and Their Priority. The Tanner Lectures on Human Values, (3), 3-87. Salt Lake City, University of Utah Press (1982).
28. –––.: The idea of an Overlapping Consensus. Oxford Journal of Legal Studies, (7), 1, 1-25 (1987).
29. –––.: Political Liberalism, revised edition. New York, Columbia University Press (1996).
30. –––.: The Domain of the Political and Overlapping Consensus. In Collected Papers, edited by S. Freeman. Cambridge Massachusetts, Harvard University Press, 473-496 (1999).
31. –––.: Justice as Fairness: A Restatement. Cambridge Massachusetts, Harvard University Press (2001).
32. Rousseau, J.J.: The Social Contract, translated by G.D.H. Cole. New York, Prometheus Books (1988).
33. The Algorithmic Justice League https://www.ajl.org Accessed on 2021/09/20