Machine morality, moral progress, and the looming environmental disaster

Running head: Machine morality and moral progress

Ben Kenward, Oxford Brookes University

Thomas Sinclair, University of Oxford

Accepted 21/04/2021 in *Cognitive Computation and Systems* for special issue *Computing Morality: Synthetic Ethical Decision Making and Behaviour*

Ethics and integrity statement: There are no relevant funding or conflicts of interest to disclose. There is no data associated with this work and therefore no relevant ethical approval.

Corresponding author: Ben Kenward, Centre for Psychological Research, Oxford Brookes University, ORCID: 0000-0002-8472-7172, bkenward@brookes.ac.uk

Abstract

The creation of artificial moral systems requires us to make difficult choices about which of varying human value sets should be instantiated. The industry-standard approach is to seek and encode moral consensus. Here we argue, based on evidence from empirical psychology, that encoding current moral consensus risks reinforcing current norms, and thus inhibiting moral progress. However, so do efforts to encode progressive norms. Machine ethics is thus caught between a rock and a hard place. The problem is particularly acute when progress beyond prevailing moral norms is particularly urgent, as is currently the case due to the inadequacy of prevailing moral norms in the face of the climate and ecological crisis.

Keywords: Machine ethics; Artificial intelligence; Morality; Moral progress; Climate Change; Ecological collapse

Machine morality

It has been argued that the development of machines programmed to make moral decisions risks atrophying human moral reasoning capacities, as humans come to outsource responsibility for decisions in which moral values are at stake to the machines [1, 2]. In this article, we identify a further concern: that moral machines risk impeding moral progress. Though our argument is general, the concern is particularly acute in the context of the urgent need for such progress at the present moment, in which prevailing moral norms are inadequate to address the ongoing environmental crises of climate breakdown [3] and ecological collapse [4] and the major threat to human life they pose.

We define a morality as a set of values, in light of which those holding the values regard some types of actions as good (to be striven after) and others as bad (to be avoided) [5]. Moral behaviour is defined as behaviour motivated by or designed to realise such values. We further define moral machines as those engaging in computational operations involving representations of values that are intended to reflect morality in order to yield moral behaviour, either on the part of the machines themselves or on the part of humans to whom the machines issue guidance. By these definitions, a fire alarm is not a moral machine, even though it tends to produce outcomes that humans approve of, because it involves no internal representations of value. By contrast, a self-driving vehicle that cannot avoid a collision but makes a decision on whether to collide with three adults or two children [6] is a moral machine because it must operate on values derived from human morality in order to make the decision. Although these definitions may be challenged in various ways, they are sufficient for our argument here.

Encoding moral values in the machines we design is increasingly a necessity to the extent that these machines are built to complete complex tasks involving interactions with humans but without human supervision, as in the case of self-driving cars and other autonomous robots. Engineers are usually regarded as obliged to programme moral behaviour in such machines [2, 7-10]. But moral machines may also be designed for decision-support roles, where it is argued that morally salient decisions that are currently normally made by humans can be made better or more efficiently by artificial systems [2, 8, 9, 11]. For example, longer prison sentences are thought to be appropriate in the case of criminals who are more likely to reoffend, but algorithms are (in some cases) better than human judges at predicting reoffending rates [12, 13]. In more private contexts, it has been argued that artificial systems could assist humans in living up to moral standards by providing moral advice [14-18].

A key question facing engineers building moral machines is this: which morality should they program? Which moral values in particular should be reflected in the machines' operations? Given the diversity of human moralities, the answer is not obvious. One approach is implied by the definition of machine ethics offered by the relevant international industrial body, according to which ethical machines are those whose behaviour is based on human moral values taken to be universal, such as human rights [2, 9]. This suggests that moral machine engineers are responding to the problem of diverse human moralities by attempting to identify universal values and encode these (though some more nuanced approaches have also been proposed [6, 19-22]). However, universal values may be harder to identify than it may seem. Even human rights are not uncontroversial [21, 23]. Similarly, the consensus on the importance of fairness, another value that is emphasised in many of the available sets of ethical guidelines for artificial intelligence, is stronger amongst demographics typical for engineers than for other humans [24]. Perhaps more importantly, even if minimal principles such as human rights were universally accepted, such principles are not sufficiently comprehensive or specific to guide all the decisions that machine ethicists want machines to be able to make [6]. No consensus account of human rights determines how a self-driving car should choose in a conflict between protecting the life of its passenger and protecting the life of a pedestrian, for example, or how an AI system advising on the distribution of scarce medical resources should weigh the lives of different people with different diseases.

Faced with these challenges, as well as the distinct concern that the systems they build might not be trusted by humans, engineers have aimed to design machines that instantiate as much as possible the values of their primary users [9, 24, 25]. Indeed, research on human trust of artificial decisions indicates that reliability, defined as the consistent production of expected decisions, is key [26]. Such research has focused primarily on perception of competence to make non-moral judgments. But a small body of work on human responses to artificial systems that are explicitly presented as making moral decisions confirms that humans tend to prefer such systems when they make judgments in line with their own judgments and those of other humans [27-29]. Further, trust in a given system tends to increase over time when people experience judgments in alignment with their expectations [26]. Especially given current low levels of trust in complex automated systems [30], it is unsurprising that organisations hoping for market success embed values in their machines that they take to reflect those of users.

Machine morality and moral progress

Most moral machines designed to perform complex tasks or provide decision-making assistance to humans will, then, reflect answers to the key question that go beyond consensus-reflecting commitments to minimal moral values such as human rights. This obviously raises concerns about the legitimacy of any particular machine morality with respect to the people with whom a given machine will interact or whose lives will be affected by the decisions it informs. These concerns are heightened by recent studies suggesting that algorithms trained or programmed to replicate human decision-making may instantiate existing human biases, effectively reflecting the distinctive moral failings of a particular group as well as the values it professes [31, 32]. Job applicant screening algorithms have been shown to discriminate against women, for example, reflecting a sexism that is at odds with the professed values of those whose behaviour the algorithm was designed to replicate [33].

However, our focus here is on the risk not that moral machines will perpetuate commonly occurring violations of prevailing moral norms, but that they will calcify those norms and the values they reflect, impeding the ordinary mechanisms of moral progress. If it were possible to restrict machine morality to minimal, universal values such as human rights, this might not seem a particularly grave concern (though we note, again, that human rights are not uncontroversial). But, as we argued above, no such restriction is realistic for the kinds of moral machines that are now being built. These machines reflect values that go beyond the minimal and universal. Even at the abstract level, there is reason to think that progress beyond these values may be desirable. If moral machines risk calcifying them, that should be cause for concern in itself. Moreover, as we will go on to argue, there is reason to think that progress beyond the values likely to be calcified by the moral machines being designed and built today is urgently needed. The cause for concern is correspondingly greater.

Before we turn to the way in which moral machines risk impeding mechanisms of moral progress, we should explain what we mean by 'moral progress'. We have said that it is a challenge for engineers to select any particular morality from among the diversity of human moralities, and we have argued that minimal moral values do not provide a sufficiently comprehensive basis for overcoming that challenge. It might be thought that we face more or less the same problem in identifying values on the basis of which to formulate a criterion of moral progress, on the basis of which to argue that moral machines risk impeding it. If we could overcome that problem by identifying the necessary values, then so, it would seem, could the engineers of moral machines.

In response: we do not rest our argument on a complete set of values on the basis of which to formulate criteria of moral progress. If we could plausibly identify any such values, that would indeed appear to answer the key question and undercut the concern about value calcification. Instead, we suggest that the following minimal principle of moral progress is plausible enough to justify seriously the concerns we raise:

> *Minimal Principle of Moral Progress*
> Moral progress obtains if the values internal to a morality change in ways that alter participant behaviour in ways that substantially reduce the likelihood of catastrophic, large-scale suffering without a corresponding increase in human rights violations.

The intuitive idea underlying the Minimal Principle is that all other things equal, a development in a set of values represents an improvement if the community whose behaviour it regulates is steered away from catastrophic outcomes such as nuclear war or severe environmental collapse. No doubt refinements could be made to the Minimal Principle that would improve the degree to which it captures this underlying idea, but we take it that the point is clear. Note that neither the Minimal Principle as we have formulated nor any principle likely to capture the underlying idea implies any determinate set of progressive values that might provide the basis for an answer to the key question facing moral machine engineers. In particular, the Minimal Principle does not imply act utilitarian values, understood as those favouring individual acts that tend to maximise aggregate utility or welfare and disfavouring individual acts that tend to do anything else [34], since such values do not provide clear protection against human rights violations [35, 36]. Note too that the Minimal Principle offers a sufficient condition of moral progress, not a necessary condition. We do not deny that there can be moral progress that the Minimal Principle does not provide grounds for classifying as such.

Machine morality is likely to impede moral progress

We have argued that because people prefer machines whose behaviour chimes with their own preferences, then machines that make moral decisions will tend to be designed to reflect the dominant behavioural moral status quo. When machines reflect the status quo, then to demonstrate that they will impede moral progress, it is sufficient to demonstrate that people will sometimes be influenced by these machines when they might instead have been influenced by progressive human voices. (For the case that moral machines do not reflect the status quo, see the next section.)

Individuals already frequently respond to moral choices by relying on external expertise such as religious authorities. In medical dilemmas, for example, even non-religious Jews sometimes ask rabbis to make decisions on their behalf in a manner that has been described as moral "outsourcing" [37, 38]. Similarly, automatic intelligent systems are valued by humans because they save us the effort and worry of making decisions for ourselves. People can easily come to heavily rely on complex automated systems, to the extent that their own motivation and ability to make relevant decisions is reduced [1, 2, 39]. A well-known example of this is the phenomenon of transportation accidents due to over-reliance on auto-pilots [40]. When automatic systems consistently make decisions that appear correct, human motivation and ability to question the systems are reduced.

It has been suggested that over-reliance on algorithmic systems in the context of migration (as used in Canada) already represents a form of moral outsourcing that is reinforcing existing norms about the values of different types of migrants [41, 42]. As the use of such systems becomes the norm, they have the potential to become the type of established cultural institution that places inertia on cultural evolution (as the sociologist Bourdieu discussed [43]).

Legal scholars have discussed the way in which formalising social rules can result in the perpetuation of practices which later generations come to regard as suboptimal but are difficult to change because of inertia within the system. Deakin [44] gives examples of "frozen accidents" in English employment law, where laws once regarded as appropriate to govern the "master-servant" relationship are no longer optimal for governing the modern employer-employee relationship but are still in place. Deakin argues this is in part because of a tendency for new legal frameworks to rely on the older ones in ways that make older ones hard to change, and in part because of a tendency to rationalise to defend the status quo (a well-known and wide-spread psychological phenomenon [45]). Given the functional similarity between machine morality and cultural institutions such as the law – they are both systems intended to issue relatively objective, reliable, and definitive judgements as to appropriate behaviour – it is reasonable to assume machine morality could also lead to similar "frozen accidents" because of similar institutional and psychological processes.

Perhaps the strongest motivation for becoming comfortable with moral outsourcing to machines is that it offers a further psychological defence against the aversive state of cognitive dissonance that is experienced when our actions are not congruent with our explicit values [46, 47]. Moral behaviour is an expression of a combination of implicit and explicit values. Implicit values are strongly internalised and influence behaviour without the necessity for conscious thought. Explicit values are those which we claim to abide by. Because talk is cheap, our explicit values tend to be more progressive [48]. Morally relevant behaviour is strongly influenced by implicitly held values, however [49-51]. For example, treatment of outgroups is at least as strongly influenced by implicitly held attitudes as by explicit attitudes [52, for a large-scale meta-analysis see 53]. Reference to moral advice from machines, if is based on existing behavioural norms (or encoded reflections of them such as laws), may therefore help us to justify behaviour that is not in line with our explicitly claimed values. Consultation with such moral machines has the potential to fulfil the same moral function as token moral efforts such as "ethical" consumption, which tends to make us feel good in a way which licenses negative behaviour even outweighing the token positive acts [54, 55].

The history of moral progress is the history of hard-fought struggles by those who succeed in persuading the majority to adopt an initially minority view [56-58]. These social movements successfully utilise a broad diversity of inter-human tactics, from reasoned persuasion to example-setting to emotive acts of self-sacrifice [59, 60]. To the extent that outsourcing morality to machines places some decision making outside the realms of these processes, rendering us less susceptible to influence from progressive human values, moral machines represent a risk to moral progress.

Why encoding progressive values is a problematic solution

It might be objected that moral machine engineers can, in recognition of the concerns we have just identified, simply programme the machines they build with more progressive values, and thereby address the concerns [14-18]. However, this is a problematic proposal. One reason for this is psychological and empirical; the other is political and philosophical.

The political and philosophical reason is as follows. As we noted above in our discussion of the Minimal Principle of Moral Progress, no determinate set of progressive moral values is implied by that principle. By contrast, any particular choice of progressive moral values, even if a morality instantiating these would indeed constitute progress according to the Minimal Principle, faces problems of legitimacy. Efforts to reflect prevailing moral norms, even if any particular choice inevitably involves privileging some contemporary moralities over others, can at least claim a sort of democratic legitimacy in virtue of that effort. By contrast, an engineer's choice of one from among the many possible sets of progressive moral values that would constitute progress according to the Minimal Principle must inevitably reflect that engineer's particular outlook and reasoning, or at best the outlook and reasoning of the select group involved in making the choice. In brief, there is no comparably legitimate substitute for the mechanisms of human moral progress that such a choice would short-circuit.

The psychological and empirical reason for thinking that the concerns we have raised cannot easily be addressed simply by programming machines with more progressive values is that machines programmed with such values risk being rejected by the humans with whom they interact. There are reasons to expect that humans will react negatively in interactions with such machines even if they agree with the idea in principle. The first has to do with the phenomenon of 'do-gooder derogation': individuals tend to react negatively to those who they perceive as promoting exemplary moral behaviour. Do-gooder derogation is in large part motivated by the desire to maintain a positive self-image in the face of potential moral criticism from others, which can be achieved by discounting the value of others and their opinions [61]. Circumstances tending to prompt do-gooder derogation include: (1) that the derogator specifically desires to act in a way that is contrary to moral advice [62, 63]; (2) that the derogator perceives the other as deliberately occupying a moral high ground [64]; and (3) that the derogator perceives the other as acting hypocritically [65]. We are unaware of any research which directly examines how these effects might apply in the case of decisions taken or advice given by moral machines. However, we note that machines designed to reflect progressive moral values may well be perceived as occupying the moral high ground by design, will offer advice which humans are tempted not to follow, and may (as machines) not be subject to their own prescriptions. It is therefore reasonable to predict that machines offering moral advice based on values more advanced than those of their advisees are subject to serious risk of do-gooder derogation.

The second reason to expect that humans will react negatively to machines programmed with progressive values is the phenomenon of 'outgroup derogation', the most relevant aspect of which involves individuals discounting the attitudes of others whom they perceive as having different social identities to themselves [66, 67]. Non-human agents are in some sense the ultimate outgroup, and indeed humans display strong negative responses towards non-human agents when these appear or behave in ways that are similar but not indistinguishable from human appearances and behaviours (the uncanny valley phenomenon [68]). Interaction with non-human systems that act in some ways similarly to humans (by offering moral advice) may therefore activate processes of outgroup derogation. Outgroup derogation can be so strong that it creates a boomerang effect, whereby exposure to an argument from a mistrusted outgroup (e.g. one's political opponents) actually reinforces the subject's contrasting opinion [69, 70]. This effect is mediated by a sense of being threatened by proponents of the opposing opinion [71]. In this context, it is worth recalling that half of US citizens fear artificial intelligence [30].

Humans may be prone to a particularly strong sense of injustice when their actions are judged by machines. One study demonstrated the curious result that when humans judged mistakes made by artificial agents, perceived injustice was a stronger predictor of mistrust of the system than was perceived harm [72]. Compare also the recent UK public protests in response to A-level school grades being generated by an algorithm (dubbed 'mutant' by the Prime Minister in response [73]).

Because machines are not human and are unlikely to have existences that humans can relate to in relevant contexts, they will be unable to engage in many of the methods of moral persuasion that, as we argued above, are a crucial mechanism of moral progress and may serve to mitigate against these kinds of backfiring reactions. Machines will be restricted to the least effective methods (such as offering sensible reasons for actions [74]). There is, therefore, little to suggest that morally progressive machines represent the solution to the concern we have raised about calcification of prevailing norms.

One reason why the problem is acute: moral psychology, the environment, and the urgent need for moral progress

To this point, we have raised our concern in relatively general terms. Moral machines may hinder the mechanisms of moral progress, and that generates the danger of progress-inhibiting value calcification. We have also argued that the politics and psychology of human reactions to artificial intelligence precludes any simple solution to this problem.

This way of raising the concern may make the dangers we have identified seem abstract and remote. This impression is mistaken, however. We conclude by providing a concrete illustration of the problem that shows it to be an acute and urgent concern.

An overwhelming body of evidence indicates that humanity is currently doing enormous damage to the Earth's climate and biodiversity [3, 4]. If we do not alter our course, human activity is likely to bring about catastrophic increases in global mean surface temperatures and devastating losses of ecosystem services and species within centuries or even decades, implying, for example, serious impacts on global food supplies [75].

Moreover, we are running out of time to make the necessary changes. In 2015, in recognition of some of these risks, governments around the world agreed to try to limit global mean surface temperature increases to 1.5°C above pre-industrial levels. A 2018 report by the Intergovernmental Panel on Climate Change estimated that for even a 50% chance of achieving that target, global carbon emissions must be cut by around half within the next ten years and to 'net zero' by 2050. Moreover, that figure assumes the availability soon of carbon capture and storage technologies that have yet to be shown to work at the necessary scale. As the IPCC claimed, this will require radical transformations in energy, industry, transport, building, agriculture, forestry, and land use. Meanwhile, as a result of our activities, the sixth mass extinction is already under way [76-78].

This parlous state indicates a clear need for moral progress, according to our minimalist definition. Further, there are reasons to think the environmental crisis might be especially subject to the risks we have associated with machine morality. Some properties of human psychology make it particularly challenging to grasp the full implications of the crisis and thus also the extent of change necessary [79]. For example, problems that are seen as distant in time and space [80, 81] and not caused by malign intent [79, 82] tend not to be perceived as moral concerns requiring urgent attention. Collective decision making currently reflects these psychological biases, and it is therefore particularly likely that machine morality will reflect the inadequate status quo.

Many populations do now espouse high levels of environmental concern – for example most Europeans explicitly claim to place a very high value on caring for the environment, caring no less than they claim to care for other people [83]. However, the lack of action (in proportion to the crisis) indicates a disconnect between explicitly espoused values and more strongly internalised implicit values that drive behaviour [84-86]. In part because of this potentially disconcerting value-action gap, individuals are particularly prone to token environmental behaviours, whereby they feel that because they have done "something", they feel they have done "enough" [54]. This frequently leads to actions that feel environmentally good having an overall negative effect because they license compensatory negative behaviour [55, 87]. This implies that consultation with machines said to be environmentally moral, but which in fact implement a status quo or insufficiently progressive morality, could be particularly counterproductive in reinforcing a false sense of moral adequacy. The prevalence of corporate greenwashing [88] already demonstrates how companies target consumers' implicit-explicit value gap, with marketing targeting explicit environmental values but products themselves targeting better internalised desires for low-cost products with little consideration for the environment [89]. It seems reasonable to assume that some commercial moral machines might do the same.

Outside the mainstream cultures of developed Western societies, environmentalist values are sometimes much more deeply embedded within cultural norms [90]. Studies in the field of cultural evolution illustrate that cultural change to reduce environmental impact is psychologically realistic: evolved cultural practices can maintain environmental impact at sustainable levels [91], and rapid moral evolution is in principle possible [92-94]. And environmentally progressive values might also be encoded in machines, of course. We note, however, that because the environment is an area where progressive values are already often espoused, it is well known that there are strong psychological defence mechanisms that operate to ignore or derogate such values when they require sacrificial action, even in cases where the individual themselves espouses the values [46, 47, 95]. Only a limited set of progressive message types can bypass these defence mechanisms to induce moral progress through properly internalised values – for example, role-modelling by prominent ingroup members at a grass-roots level [96]. As outlined above, these are the types of message which machines would be very challenged to deliver.

In summary, evidence indicates that although the environmental values which most people display are generally insufficient to motivate necessary environmentalist behaviour [97], this could change. A shift away from egoistic individual rights-based norms, towards more strongly internalised environmentalist duty-based norms, might well result in increased environmentalist behaviour and therefore represent much-needed moral progress. However, the risks of stagnation or backlash associated with machine morality indicate that it may be particularly ill-suited to this task.

Conclusion

We have argued that powerful incentives structure the choices of engineers concerning which moral values the machines they build should reflect, as a result of which the machines they build will tend to reflect prevailing moral values. We also argued that mechanisms of moral progress are likely to be inhibited by moral machines. The upshot of these two arguments is that moral machines present a real risk of calcifying prevailing moral values. Moreover, the risk we have identified is not a mere abstract, theoretical possibility. On the contrary, as humans face the catastrophic dangers of climate change and ecological destruction as a result of their ongoing engagement in activities that they take to be licensed by prevailing moral norms, the risk of value calcification associated with moral machines is clear, acute, and urgent.

However, we also acknowledge that it is difficult to assess how great this risk is, because there are currently few machines making moral decisions, and thus little directly relevant data. Although there is no clear reason to believe that encoding the status quo could assist with moral progress, and reasons to believe moral machines will indeed encode the status quo, we also note that there are arguments that moral progress might be assisted by moral machines that are progressive [14-18]. Further, we agree that artificial intelligence in general might prove crucial in supporting decisions of moral relevance, for example regarding the environment [98, 99]. However, in line with previous critics of machine ethics [100-103], we suggest that automatic systems can be constructed for human decision support without computing moral values.

Our aim has been to identify and articulate a risk that has hitherto gone largely unrecognised. For reasons of space and the limits of available evidence, we have not offered an assessment as to the gravity of the risk either in general or in the present moment or proposed solutions to it. However, we believe we have said enough to warrant the inclusion of this risk alongside other, more familiar risks associated with AI and related technologies in the thinking of those who consider the creation of moral machines.

References

1. Danaher J. The rise of the robots and the crisis of moral patiency. AI & SOCIETY. 2019;34(1):129-36.

2. Cave S, Nyrup R, Vold K, Weller A. Motivations and Risks of Machine Ethics. Proceedings of the IEEE. 2019;107(3):562-74.

3. IPCC. Summary for policymakers. In: Masson-Delmotte V, Zhai P, Pörtner HO, Roberts D, Skea J, P. R. Shukla, et al., editors. Global warming of 15°C An IPCC

Special Report on the impacts of global warming of 15°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty Geneva, Switzerland: World Meteorological Organization; 2018. p. 32.

4. IPBES. Summary for policymakers. In: Díaz S, Settele J, E.S. ESB, Ngo HT, Guèze M, Agard J, et al., editors. The global assessment report on biodiversity and ecosystem services. IPBES secretariat: Bonn, Germany; 2019. p. 56.

5. Gert B, Gert J. The Definition of Morality. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy (Fall 2020 Edition)2020.

6. Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. The Moral Machine experiment. Nature. 2018;563(7729):59-64.

7. Simpson TW, Müller VC. Just war and robots' killings. The Philosophical Quarterly. 2016;66(263):302-22.

8. Malle BF. Integrating robot ethics and machine morality: the study and design of moral competence in robots. Ethics and Information Technology. 2016;18:243-56.

9. IEEE. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: 2019. Available from: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

10. Bryson JJ. Robots should be slaves. In: Wilks Y, editor. Close Engagements with Artificial Companions: John Benjamins; 2010. p. 63-74.

11. Cervantes J-A, López S, Rodríguez L-F, Cervantes S, Cervantes F, Ramos F. Artificial Moral Agents: A Survey of the Current Status. Science and Engineering Ethics. 2020;26(2):501-32.

12. Lin ZJ, Jung J, Goel S, Skeem J. The limits of human predictions of recidivism. Science Advances. 2020;6(7):eaaz0652.

13. Kehl DL, Kessler SA. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. Responsive Communities Initiative, Berkman Klein Center for Internet & Society, Harvard Law School 2017 [Available from: http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041.

14. Savulescu J, Maslen H. Moral enhancement and artificial intelligence: Moral AI? Beyond Artificial Intelligence: Springer; 2015. p. 79-95.

15. Klincewicz M. Artificial intelligence as a means to moral enhancement. Studies in Logic, Grammar and Rhetoric. 2016;48(1):171-87.

16. Giubilini A, Savulescu J. The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence. Philosophy & Technology. 2018;31(2):169-88.

17. Anderson M, Anderson SL. Introduction to Part V. In: Anderson M, Anderson SL, editors. Machine Ethics. Cambridge: Cambridge University Press; 2011. p. 495-8.

18. Seville H, Field D. What can AI do for ethics. AISB QUARTERLY. 2000:31-4.

19. Gordon J-S. Building Moral Robots: Ethical Pitfalls and Challenges. Science and Engineering Ethics. 2020;26(1):141-57.

20. Santos-Lang C. Our responsibility to manage evaluative diversity. SIGCAS Comput Soc. 2014;44(2):16–9.

21. Wong P-H. Cultural Differences as Excuses? Human Rights and Cultural Values in Global Ethics and Governance of AI. Philosophy & Technology. 2020;33(4):705-15.

22. Ecoffet A, Lehman J. Reinforcement Learning Under Moral Uncertainty. arXiv:200604734 preprint. 2020.

23. Deacon R. Human Rights as Imperialism. Theoria. 2003;50(102):126-38.

24. Hagendorff T. The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines. 2020;30(1):99-120.

25. IBM. Building trust in AI 2019 [Available from: https://www.ibm.com/watson/advantage-reports/future-of-artificial-intelligence/building-trust-in-ai.html.

26. Glikson E, Woolley AW. Human Trust in Artificial Intelligence: Review of Empirical Research. Academy of Management Annals. 2020;14(2):627-60.

27. Yokoi R, Nakayachi K. Trust in autonomous cars: exploring the role of shared moral values, reasoning, and emotion in safety-critical decisions. Human factors. 2020:0018720820933041.

28. Malle BF, Scheutz M, Forlizzi J, Voiklis J, editors. Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI); 2016 7-10 March 2016.

29. Banks J. Good Robots, Bad Robots: Morally Valenced Behavior Effects on Perceived Mind, Morality, and Trust. International Journal of Social Robotics. 2020.

30. Blumberg Capital. Artificial Intelligence in 2019: Getting past the adoption tipping point 2019 [Available from: https://www.blumbergcapital.com/ai-in-2019/.

31. Criado N, Such JM. Digital Discrimination. In: Yeung K, Lodge M, editors. Algorithmic Regulation. Oxford: Oxford University Press; 2019.

32. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017;356(6334):183.

33. Saka E. Big Data and Gender - Biased Algorithms. The International Encyclopedia of Gender, Media, and Communication. 2020:1-4.

34. Smart JJC. An outline of a utilitarian system of ethics. In: Smart JJC, Williams B, editors. Utilitarianism: For and against: Cambridge University Press; 1973.

35. Sumner LW. Mill’s Theory of Rights. In: West HR, editor. The Blackwell Guide to Mill’s Utilitarianism: Blackwell; 2006.

36. Lyons D. Mill‘s Theory of Justice. In: Goldman AI, Kim J, editors. Values and Morals: Essays in Honor of William Frankena, Charles Stevenson, and Richard Brandt: Springer; 1978.

37. Ivry T, Teman E. Shouldering Moral Responsibility: The Division of Moral Labor among Pregnant Women, Rabbis, and Doctors. American Anthropologist. 2019;121(4):857-69.

38. Keshet Y, Liberman I. Coping with Illness and Threat: Why Non-religious Jews Choose to Consult Rabbis on Healthcare Issues. Journal of Religion and Health. 2014;53(4):1146-60.

39. Reason J. Human error. Cambridge: Cambridge University Press; 1990.

40. Harford T. Crash: How Computers are Setting us up for Disaster. The Guardian. 2016.

41. Beduschi A. International migration management in the age of artificial intelligence. Migration Studies. 2020.

42. Molnar P, Gill L. Bots at the Gate: a human rights analysis of automated decision-making in Canada's immigration and refugee system. University of Toronto. 2018.

43. Wacquant L. Habitus. In: Beckert J, Zafirovski M, editors. International Encyclopedia of Economic Sociology. London: Routledge; 2004. p. 315–9.

44. Deakin S. Evolution for Our Time: A Theory of Legal Memetics. Current Legal Problems. 2002;55(1):1-42.

45. Laurin K. Inaugurating Rationalization: Three Field Studies Find Increased Rationalization When Anticipated Realities Become Current. Psychological Science. 2018;29(4):483-95.

46. Norgaard KM. "People Want to Protect Themselves a Little Bit": Emotions, Denial, and Social Movement Nonparticipation. Sociological Inquiry. 2006;76(3):372-96.

47. Adams M. Ecological crisis, sustainability and the psychosocial subject: Springer; 2016.

48. FeldmanHall O, Mobbs D, Evans D, Hiscox L, Navrady L, Dalgleish T. What we say and what we do: The relationship between real and hypothetical moral choices. Cognition. 2012;123(3):434-41.

49. Haidt J. Moral psychology for the twenty-first century. Journal of Moral Education. 2013;42(3):281-97.

50. Cameron CD, Payne BK, Sinnott-Armstrong W, Scheffer JA, Inzlicht M. Implicit moral evaluations: A multinomial modeling approach. Cognition. 2017;158:224-41.

51. Hofmann W, Baumert A. Immediate affect as a basis for intuitive moral judgement: An adaptation of the affect misattribution procedure. Cognition and Emotion. 2010;24(3):522-35.

52. Jost JT. The IAT Is Dead, Long Live the IAT: Context-Sensitive Measures of Implicit Attitudes Are Indispensable to Social and Political Psychology. Current Directions in Psychological Science. 2018;28(1):10-9.

53. Kurdi B, Seitchik AE, Axt JR, Carroll TJ, Karapetyan A, Kaushik N, et al. Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. American psychologist. 2019;74(5):569.

54. Sachdeva S, Jordan J, Mazar N. Green consumerism: moral motivations to a sustainable future. Current Opinion in Psychology. 2015;6:60-5.

55. Sörqvist P, Langeborg L. Why People Harm the Environment Although They Try to Treat It Well: An Evolutionary-Cognitive Perspective on Climate Compensation. Frontiers in Psychology. 2019;10(348).

56. Moody-Adams MM. Moral Progress and Human Agency. Ethical Theory and Moral Practice. 2017;20(1):153-68.

57. Anderson E. Social movements, experiments in living, and moral Progress: Case studies from Britain's abolition of slavery. The Lindley Lecture, University of Kansas, Department of Philosophy. 2014.

58. Jamieson D. Slavery, Carbon, and Moral Progress. Ethical Theory and Moral Practice. 2017;20(1):169-83.

59. Perloff RM. The dynamics of persuasion: Communication and attitudes in the twenty-first century: Routledge; 2020.

60. Smithey LA. Social movement strategy, tactics, and collective identity 1. Sociology Compass. 2009;3(4):658-71.

61. Minson JA, Monin B. Do-Gooder Derogation: Disparaging Morally Motivated Minorities to Defuse Anticipated Reproach. Social Psychological and Personality Science. 2011;3(2):200-7.

62. Bolderdijk JW, Brouwer C, Cornelissen G. When Do Morally Motivated Innovators Elicit Inspiration Instead of Irritation? Frontiers in Psychology. 2018;8(2362).

63. Zane DM, Irwin JR, Reczek RW. Do less ethical consumers denigrate more ethical consumers? The effect of willful ignorance on judgments of others. Journal of Consumer Psychology. 2016;26(3):337-49.

64. Kurz T, Prosser AMB, Rabinovich A, O'Neill S. Could Vegans and Lycra Cyclists be Bad for the Planet? Theorizing the Role of Moralized Minority Practice Identities in Processes of Societal-Level Change. Journal of Social Issues. 2020;76(1):86-100.

65. Sparkman G, Attari SZ. Credibility, communication, and climate change: How lifestyle inconsistency and do-gooder derogation impact decarbonization advocacy. Energy Research & Social Science. 2020;59:101290.

66. Hogg MA, Smith JR. Attitudes in social context: A social identity perspective. European Review of Social Psychology. 2007;18(1):89-131.

67. Mackie DM, Gastardo-Conaco MC, Skelly JJ. Knowledge of the advocated position and the processing of in-group and out-group persuasive messages. Personality and Social Psychology Bulletin. 1992;18(2):145-51.

68. Syrdal DS, Dautenhahn K, Koay KL, Walters ML. The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. Adaptive and emergent behaviour and complex systems. 2009.

69. Hart PS, Nisbet EC. Boomerang Effects in Science Communication: How Motivated Reasoning and Identity Cues Amplify Opinion Polarization About Climate Mitigation Policies. Communication Research. 2011;39(6):701-23.

70. Ma Y, Dixon G, Hmielowski JD. Psychological Reactance From Reading Basic Facts on Climate Change: The Role of Prior Views and Political Identification. Environmental Communication. 2019;13(1):71-86.

71. Hoffarth MR, Hodson G. Green on the outside, red on the inside: Perceived environmentalist threat as a factor explaining political polarization of climate change. Journal of Environmental Psychology. 2016;45:40-9.

72. Sullivan Y, de Bourmont M, Dunaway M. Appraisals of harms and injustice trigger an eerie feeling that decreases trust in artificial intelligence systems. Annals of Operations Research. 2020.

73. Paulden T. A cutting re-mark. Significance. 2020;17(5):4-5.

74. Blair JA. The persuasive ineffectiveness of arguing and arguments. OSSA Conference Archive. 2020;10.

75. FAO. The State of the World’s Biodiversity for Food and Agriculture, J. Bélanger & D. Pilling (eds.). FAO Commission on Genetic Resources for Food and Agriculture Assessments. Rome: http://www.fao.org/3/CA3129EN/CA3129EN.pdf; 2019.

76. Ceballos G, Ehrlich PR, Dirzo R. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. Proceedings of the national academy of sciences. 2017;114(30):E6089-E96.

77. Ceballos G, Ehrlich PR, Barnosky AD, García A, Pringle RM, Palmer TM. Accelerated modern human–induced species losses: Entering the sixth mass extinction. Science advances. 2015;1(5):e1400253.

78. Barnosky AD, Matzke N, Tomiya S, Wogan GO, Swartz B, Quental TB, et al. Has the Earth's sixth mass extinction already arrived? Nature. 2011;471(7336):51-7.

79. Markowitz EM, Shariff AF. Climate change and moral judgement. Nature Climate Change. 2012;2(4):243-7.

80. Singh AS, Zwickle A, Bruskotter JT, Wilson R. The perceived psychological distance of climate change impacts and its influence on support for adaptation policy. Environmental Science & Policy. 2017;73:93-9.

81. Spence A, Poortinga W, Pidgeon N. The Psychological Distance of Climate Change. Risk Analysis. 2012;32(6):957-72.

82. Greene JD, Cushman FA, Stewart LE, Lowenberg K, Nystrom LE, Cohen JD. Pushing moral buttons: The interaction between personal force and intention in moral judgment. Cognition. 2009;111(3):364-71.

83. Bouman T, Steg L. Motivating Society-wide Pro-environmental Change. One Earth. 2019;1(1):27-30.

84. Farjam M, Nikolaychuk O, Bravo G. Experimental evidence of an environmental attitude-behavior gap in high-cost situations. Ecological Economics. 2019;166:106434.

85. Beattie G, McGuire L. Consumption and climate change: Why we say one thing but do another in the face of our greatest threat. Semiotica. 2016;2016(213):493-538.

86. Thomas GO, Walker I. The Development and Validation of an Implicit Measure Based on Biospheric Values. Environment and Behavior. 2014;48(5):659-85.

87. Maki A, Carrico AR, Raimi KT, Truelove HB, Araujo B, Yeung KL. Meta-analysis of pro-environmental behaviour spillover. Nature Sustainability. 2019;2(4):307-15.

88. Lyon TP, Montgomery AW. The Means and End of Greenwash. Organization & Environment. 2015;28(2):223-49.

89. van 't Veld K. Eco-Labels: Modeling the Consumer Side. Annual Review of Resource Economics. 2020;12(1):187-207.

90. Gratani M, Sutton SG, Butler JRA, Bohensky EL, Foale S. Indigenous environmental values as human values. Cogent Social Sciences. 2016;2(1):1185811.

91. Brooks JS, Waring TM, Borgerhoff Mulder M, Richerson PJ. Applying cultural evolution to sustainability challenges: an introduction to the special issue. Sustainability Science. 2018;13(1):1-8.

92. Lindström B, Jangard S, Selbing I, Olsson A. The role of a "common is moral" heuristic in the stability and change of moral norms. Journal of Experimental Psychology: General. 2018;147(2):228-42.

93. Fogarty L, Kandler A. The fundamentals of cultural adaptation: implications for human adaptation. Scientific Reports. 2020;10(1):14318.

94. Wheeler MA, McGrath MJ, Haslam N. Twentieth century morality: The rise and fall of moral concepts from 1900 to 2007. PLOS ONE. 2019;14(2):e0212267.

95. Lamb WF, Mattioli G, Levi S, Roberts JT, Capstick S, Creutzig F, et al. Discourses of climate delay. Global Sustainability. 2020;3:e17.

96. Grabs J, Langen N, Maschkowski G, Schäpke N. Understanding role models for change: a multilevel analysis of success factors of grassroots initiatives for sustainable consumption. Journal of Cleaner Production. 2016;134:98-111.

97. Persson I, Savulescu J. Unfit for the future: the need for moral enhancement: OUP Oxford; 2012.

98. Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, et al. The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Communications. 2020;11(1):233.

99. Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, et al. Tackling climate change with machine learning. arXiv:190605433 preprint. 2019.

100. van Wynsberghe A, Robbins S. Critiquing the Reasons for Making Artificial Moral Agents. Science and Engineering Ethics. 2019;25(3):719-35.

101. Bryson JJ. Patiency is not a virtue: the design of intelligent systems and systems of ethics. Ethics and Information Technology. 2018;20(1):15-26.

102. Ryan M. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. Science and Engineering Ethics. 2020;26(5):2749-67.

103. Nyrup R, Whittlestone J. Why Value Judgements Should Not Be Automated 2019 [Available from: https://doi.org/10.17863/CAM.41552.