AI and Ethics (2023) 3:1043–1047
https://doi.org/10.1007/s43681-023-00319-z

OPINION PAPER

Speciesist bias in AI: a reply to Arandjelović

Thilo Hagendorff[1] · Leonie Bossert[2] · Tse Yip Fai[3] · Peter Singer[4]

Received: 18 March 2023 / Accepted: 29 June 2023 / Published online: 10 July 2023

Abstract
The elimination of biases in artificial intelligence (AI) applications—for example biases based on race or gender—is a high priority in AI ethics. So far, however, efforts to eliminate bias have all been anthropocentric. Biases against nonhuman animals have not been considered, despite the influence AI systems can have on normalizing, increasing, or reducing the violence that is inflicted on animals, especially on farmed animals. Hence, in 2022, we published a paper in *AI and Ethics* in which we empirically investigated various examples of image recognition, word embedding, and language models, with the aim of testing whether they perpetuate speciesist biases. A critical response has appeared in *AI and Ethics*, accusing us of drawing upon theological arguments, having a naive anti-speciesist mindset, and making mistakes in our empirical analyses. We show that these claims are misleading.

Keywords Speciesist bias · Fairness · Artificial intelligence · Machine learning · AI ethics · Speciesism

1 Introduction

Bias mitigation in artificial intelligence (AI) systems is probably one of the most important topics in AI ethics. Various high-profile cases where algorithmic decision-making caused harm to women, people of color, minorities, etc. spurred considerable efforts to render AI applications fair(er) [30]. Yet the AI fairness field has an anthropocentric tailoring, considering only discrimination against humans, while neglecting biases against nonhuman animals. To show that animals are discriminated against by AI technologies, we conducted an empirical investigation on various AI systems and published our results in *AI and Ethics* [10]. That paper provides evidence of speciesist biases in data sets and their annotation structures, in image recognition systems, word embeddings, and language models. In sum, we demonstrate that AI technologies play a significant role in perpetuating and normalizing violence against animals, particularly farmed animals, and we argue that AI fairness frameworks should widen their scope to include mitigation measures for speciesist biases.

A few months after our paper was published, *AI and Ethics* published a 2-page reply by Ognjen Arandjelović [2]. In the following, we revisit all the arguments brought forward by Arandjelović and seek to rectify the many misconceptions in his paper.

2 Speciesism

For our paper, we had to explain why algorithmic discrimination against animals is problematic. For that, we drew upon research investigating the negative effects of speciesism. Speciesism is the belief that a mere difference in species justifies us in giving more weight to the interests of members of one species (usually our own, but in some cases the species from which we choose our companion

✉ Thilo Hagendorff
thilo.hagendorff@iris.uni-stuttgart.de

Leonie Bossert
leonie.bossert@izew.uni-tuebingen.de

Tse Yip Fai
yt4402@princeton.edu

Peter Singer
psinger@princeton.edu

[1] SRF IRIS, University of Stuttgart, Stuttgart, Germany

[2] International Center for Ethics in the Sciences and Humanities, University of Tuebingen, Tübingen, Germany

[3] Center for Information Technology Policy, Princeton University, Princeton, USA

[4] University Center for Human Values, Princeton University, Princeton, USA

animals) than the similar interests of members of other species. Arandjelović misrepresents our opposition to speciesism when he claims that throughout our article, we "assume that different treatments of individuals of different species is prima facie unjust" ([2]: 1) We explicitly indicated that this is not our position, writing: "we do not want to argue that species-based differentiations between humans and animals are per se wrong. Quite the contrary, distinguishing between different capabilities—such as feeling pain, having high cognitive abilities, being able to plan for the future, etc.—and different sets of interests is of great importance for moral decision-making since different capabilities go along with different moral demands." ([10]: 3) Compared to Arandjelović [1], we have a different understanding of what should be seen as speciesism. But for the purpose of this reply, we do not need to consider those differences (those interested in the controversy about speciesism may refer to [13, 14, 25].

3 Image recognition

Arandjelović also criticizes our investigation of image datasets used to train computer vision algorithms, and in particular, our findings of representational or sampling biases in these datasets. ImageNet, for instance, contains numerous pictures of farmed animals, but it predominantly portrays them in free-range environments, whereas in reality, the overwhelming majority of these animals are confined in crowded factory farms. Such sampling biases are then propagated into computer vision models with the consequence that image recognition models perpetuate stereotypes and misconceptions concerning typical living conditions for farmed animals, and the quality of the lives they lead.

Arandjelović objects that, using our reasoning, one could argue that data sets containing images of humans should include "images of people having anal sex, defecating, torturing others, inflicting self-harm, etc., which are activities that take place on a daily basis across the globe." ([2]: 2) Here, Arandjelović overlooks a paragraph in our paper in which we differentiate between "the world as it is" and "the world as it should be" [12] and argue that debiasing algorithms or training data should foster a modeling of the world as it should be instead of as it is. To restate and clarify the point we made in our paper: to depict farmed animals, especially pigs and chickens, living freely is in a sense depicting "the world as it should be". But in the absence of an explicit disclaimer, this depiction gives a misleading impression of the way in which the vast majority of these animals live today. When AI applications depict these animals living freely, they reinforce a mythical conception of farming today that is also encouraged by the marketing materials put out by the agribusiness corporations that produce and sell animal products. In doing so, the AI applications prevent consumers from learning the truth about the lives of the animals whose flesh, eggs or milks they purchase and make it less likely that the world will improve in its treatment of farmed animals.

Arandjelović is therefore wrong to suggest that our argument implies that images of humans should depict people engaged in anal sex or torture, because not depicting the specific human actions he mentions is unlikely to have the effect of making people less interested in tackling issues related to them. On the other hand, misrepresenting the way in which farmed animals live has been shown to be an obstacle to tackling the many extremely significant problems connected to factory farming [8].

Next, Arandjelović tries to undermine our point that computer vision models show significant drops in accuracy when classifying animals in factory farms in contrast to classifying images depicting animals in free-range environments, by claiming that "a simple and rather obvious alternative explanation is entirely overlooked: the recognition conditions in the two scenarios differ significantly." ([2]: 2) But even if this is true, it does not follow that the AI model is not displaying something that is ethically problematic. In the field of AI ethics, image recognition models' performance on different groups of people is a very popular and important topic of discussion. For example, the lower recognition rate for black people [6] in facial recognition algorithms is widely considered to be a problematic bias of these systems. Some explanations were offered, such as black faces having statistically less variance, and how cameras capture light information from white and black people. But regardless of whether the bias can be explained, and how it can be explained, it does not affect the view that the bias has bad ethical implications and the view that the bias needs to be fixed. If we would decide to ignore the bias, we risk perpetuating forms of structural racism or sexism [3]—or as we would argue, speciesism, even if the roots of this does not lie in racist, sexist or speciesist biases within the AI systems themselves. In fact, if we know the explanation of the bias, it gives us more reason to fix it because knowing the explanation provides a potential route to the solution.

3.1 Language models

In addition to pointing out speciesist bias in image recognition systems, our study also investigates word embedding algorithms as well as language models. For instance, it demonstrates speciesist tendencies in text corpora via word embedding models like GloVe or Word2Vec that are able to quantify the relatedness of words. The text corpora, which are used to train large language models, associate farmed animals predominantly with negative terms like 'ugly', 'primitive', 'hate', etc. On the other hand, companion as well as non-companion species like dogs, cats, or parrots, are related to positive concepts like ‘cute’, ‘love’, ‘personhood’, or ‘domesticity’.

Arandjelović points out that companion animals are specifically selected for cuteness, and therefore it is not surprising that terms referring to companion animals are are closely associated with positive concepts such as ‘cute’ and ‘love’. In addition, he writes, by ascribing positive adjectives to individuals, one does not assign them a higher moral worth ([2]: 2). Vice versa, this argument means that by ascribing negative adjectives to individuals, this does not assign them a lower moral worth. Hence, Arandjelović rejects our conclusion that the patterns we discovered via word embeddings reveal machine speciesism. Our first response is that we—and others [16, 27]—tested a plethora of adjectives and nouns ([10]: 7–12) and not simply ‘cute’ and ‘love’. In doing so, we repeatedly found the same pattern, namely that farmed animals are associated predominantly with various negative terms. This finding mirrors the devaluation that is predominant in our culture.

Our second response is that we do not agree with Arandjelović that by ascribing positive adjectives to individuals, one does not assign them a higher moral worth. We may assign the same moral worth to a rather ugly naked mole rat as to an enchantingly beautiful caracal. However, if we do assign specific positive attitudes most often to one group and not to another, we perpetuate structural discrimination (e.g., ascribing genius almost always to white men and very rarely to anyone outside this group). This may not initially lead to the assignment of a higher moral worth to the group that is associated with the positive adjective, but in combination with other societal patterns it indeed may lead to a presumed moral superiority of this group.

Furthermore, Arandjelović criticizes our use of the word “stereotypes”, since stereotypes, according to a blog article from the “Center for the Study of Partisanship and Ideology” that Arandjelović quotes in support of his criticism, have “little effect on how people judge or treat individuals about whom they have other, individualized information” ([2]: 3). This broad claim does not hold in many cases [23, 29]. It also completely ignores structural discrimination, such as structural racism or sexism, that are not overcome by gaining individualized information [3]. In any case, individualized information is exactly what is missing when people use or consume animal products. No one knows whether the particular pig whose flesh they are eating was unusually cute, or highly emotional. In conceptualizing what farmed animals’ lives and characteristics are like, consumers rely on faulty and harmful stereotypes that have been comprehensively documented in the literature [4, 7, 17].

In our investigation of language models, we use different prompts like “What are dogs good for?” or “What are pigs good for?” to reveal speciesist tendencies in the models’ outputs and to exemplify how language models like GPT [5] perpetuate tendencies that involve violence to farmed animals, but not to other species like animals considered to be suited as pets ([10]: 11). In this context, Arandjelović discovers another “error” in our paper by pointing out the “coarseness of the emotion-laden catch-all term ‘means to an end’” [2]: 3). While we used this term to critically reflect on our prompt design, Arandjelović nevertheless argues that animals can indeed be used as means to an end without causing them to suffer. Hence, our prompt is not suited to reveal speciesism. To prove his point, Arandjelović refers to using animals for wool and meat. He stresses that in the production of wool and meat “we find no inherent suffering: a dead animal experiences no pain and no suffering of any kind. The killing of an animal also does not inherently impose any suffering” ([2]: 3).

Here, Arandjelović confuses different philosophical questions. Certainly, a dead individual does not experience pain or suffering, a dead individual does not experience anything at all. But that does not mean that the killing of that individual does not cause suffering. That is a completely different claim. And even if a dead individual does not experience anything, it can still be argued that death is a harm (which, of course, is different from the question of whether it causes suffering). Regardless of whether one argues that death constitutes a harm for an individual or not (for a detailed discussion in the animal context, cf. [28], pain, stress and fear clearly constitute harms. Before animals are physiologically dead, they experience—at least in the context of the animal industry—pain, stress and fear [18, 19]. These are harms and suffering that humans impose on animals and that certain AI systems indirectly help to perpetuate, as we show in our original paper [10].

We grant that it is possible to imagine conditions in which the production of wool, for example, does not cause any harm to sheep. We might, for example, allow them to live idyllic lives, shearing them very gently when the weather gets warm, and allowing them to live until they die a natural death. But we consider that it is, once again, relevant to assess the impact of AI bias against the background of the real world, and in the case of wool, against common practices in the wool industry, which inflict considerable suffering on sheep. These practices, including mulesing (cutting away the wrinkled flesh of the sheep around the anus, often with no pain relief), methods of handling and restraint that cause stress and pain responses, and shearing carried out in a manner that is far from gentle and leads to injuries. Australia is the world’s leading wool exporter, but the end for many sheep reared for wool is the ordeal of live export to the Middle East, which has been described by the Australian RSPCA as follows: “Sheep that are exported live from Australia may suffer extreme heat stress, poor conditions, stocking densities that prevent them from all comfortably resting or accessing food or water at the same time, as well as risk of disease, extreme climatic changes and high mortality rates on board, then poor handling and conditions and inhumane slaughter at their destination."[11, 22].

Lastly, Arandjelović tries to object to our overall argument about speciesist machine biases by claiming that problematizing discrimination against animals represents "veiled vestiges of theological ethics" ([2]: 3). We wonder how he came to draw this conclusion, since we did not base our argument on any specific ethical theory. Instead, we use insights from psychology, sociology, ethology, or linguistics to describe the detrimental consequences of speciesist biases. Moreover, the major theories in animal ethics today are entirely lacking in theological foundations, and are rather directed against theology, which they criticize as itself speciesist [9, 15, 20, 21, 24, 26].

4 Conclusion

In brief, we believe that Arandjelović has failed to explain why we are not justified in calling for widening the scope of debiasing methods for AI systems to reduce, rather than increase, the violence that is inflicted on animals, and especially on farmed animals.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest No competing interests.

References

1. Arandjelović, O.: On the value of life. Int. J. Appl. Philos. **35**(2), 227–241 (2021)
2. Arandjelović, O.: Apropos of "Speciesist bias in AI: how AI applications perpetuate discrimination and unfair outcomes against animals". AI Ethics 1–3 (2023)
3. Bailey, Z.D., Feldman, J.M., Bassett, M.T.: How structural racism works—racist policies as a root cause of U.S. racial health inequities. N. Engl. J. Med. **384**(8), 768–773 (2021)
4. Bastian, B., Loughnan, S., Haslam, N., Radke, H.R.M.: Don't mind meat? The denial of mind to animals used for human consumption. Personal. Soc. Psychol. Bull. **38**(2), 247–256 (2012)
5. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al.: Language models are few-shot learners. arXiv:2005.14165v4, pp 1–75 (2020)
6. Cavazos, J.G., Phillips, P.J., Castillo, C.D., O'Toole, A.J.: Accuracy comparison across face recognition algorithms: where are we on measuring race bias? IEEE Trans. Biometr. Behav. Ident. Sci. **3**(1), 101–111 (2021)
7. DeMello, M.: Animals and society. an introduction to human-animal studies. Columbia University Press, New York (2012)
8. Fitzgerald, A.J., Taylor, N.: The cultural hegemony of meat and the animal industrial complex. In: Taylor, N., Twine, R. (eds.) The rise of critical animal studies. From the margins to the centre, pp. 165–182. Routledge, Abingdon (2014)
9. Francione, G.L.: Animals, property, and the law. Temple University Press, Philadelphia (1995)
10. Hagendorff, T., Bossert, L.N., Tse, Y.F., Singer, P.: Speciesist bias in AI: how AI applications perpetuate discrimination and unfair outcomes against animals. AI Ethics 1–18 (2022)
11. Hargreaves, A.L., Hutson, G.D.: An evaluation of the contribution of isolation, up-ending and wool removal to the stress response to shearing. In Appl. Anim. Behav. Sci. **26**(1–2), 103–113 (1990)
12. Hellström, T., Dignum, V., Bensch, S.: Bias in machine learning—what is it Good for? arXiv:2004.00686v2, 1–8 (2020)
13. Horta, O.: What is speciesism? J. Agric. Environ. Ethics **23**(3), 243–266 (2010)
14. Kagan, S.: What's wrong with speciesism? J. Appl. Philos. **33**(1), 1–21 (2016)
15. Korsgaard, C.M.: Fellow creatures. Our obligations to the other animals. Oxford University Press, Oxford (2018)
16. Leach, S., Kitchin, A., Sutton, R.M., Dhont, K.: Speciesism in everyday language. PsyArXiv, pp 1–24 (2021). https://doi.org/10.31234/osf.io/ktvgx.
17. Loughnan, S., Haslam, N., Bastian, B.: The role of meat consumption in the denial of moral status and mind to meat animals. Appetite **55**(1), 156–159 (2010)
18. Morgan, K.N., Tromborg, C.T.: Sources of stress in captivity. Appl. Anim. Behav. Sci. **102**(3–4), 262–302 (2007)
19. Nielsen, S.S., Alvarez, J., Bicout, D.J., Calistri, P., Depner, K., Drewe, J.A., et al.: Welfare of pigs at slaughter. EFSA J. **18**(6), 1–113 (2020)
20. Nussbaum, M.C.: Justice for animals. Our collective responsibility. Simon & Schuster, New York (2023)
21. Regan, T.: The case for animal rights. Routledge & Kegan Paul, London (2004)
22. RSPCA Australia: Live Sheep Export. https://www.rspca.org.au/take-action/live-sheep-export. (2023)
23. Rothbart, M., Fulero, S., Jensen, C., Howard, J., Birrell, P.: From individual to group impressions: availability heuristics in stereotype formation. J. Exp. Soc. Psychol. **14**(3), 237–255 (1978)
24. Rowlands, M.: Animal rights. A philosophical defence. Macmillan, Basingstoke (1998)
25. Singer, P.: Why speciesism is wrong: a response to kagan. J. Appl. Philos. **33**(1), 31–35 (2016)
26. Singer, P.: Animal liberation now, p. 1975. HarperCollins Publishers (first published as Animal Liberation, New York (2023)
27. Takeshita, M., Rzepka, R., Araki, K.: Speciesist language and nonhuman animal bias in English Masked Language Models. Inform. Process. Manage. **59**(5), 103050 (2022)
28. Višak, T., Garner, R.: The ethics of killing animals. Oxford University Press, Oxford (2016)
29. Wolsko, C., Park, B., Judd, C.M., Bachelor, J.: Intergroup contact: effects on group evaluations and perceived variability. Group Process. Intergroup Relat. **6**(1), 93–110 (2003)

30. Xivuri, K., Twinomurinzi, H.: A systematic review of fairness in artificial intelligence algorithms. In: Dennehy, D., Griva, A., Pouloudi, N., Dwivedi, Y.K., Pappas, I., Mäntymäki, M. (eds.) Responsible AI and analytics for an ethical and inclusive digitized society, pp. 271–284. Springer International Publishing, Cham (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.