https://doi.org/10.1007/s43681-022-00255-4

COMMENTARY

Apropos of "Speciesist bias in AI: how AI applications perpetuate discrimination and unfair outcomes against animals"

Ognjen Arandjelović[1]

Received: 2 November 2022 / Accepted: 19 December 2022

Abstract

The present comment concerns a recent *AI & Ethics* article which purports to report evidence of speciesist bias in various popular computer vision (CV) and natural language processing (NLP) machine learning models described in the literature. I examine the authors' analysis and show it, ironically, to be prejudicial, often being founded on poorly conceived assumptions and suffering from fallacious and insufficiently rigorous reasoning, its appeal in large part relying on the extant consensus in the community.

Keywords Fairness · Value of life · Exploitation · Computer vision · Machine learning

The present comment concerns the article entitled "Speciesist bias in AI: how AI applications perpetuate discrimination and unfair outcomes against animals" published online in *AI & Ethics* and authored by Hagendorff et al. [3]. While as a researcher in machine learning and computer vision I found the authors' results interesting, as a philosopher I found the interpretation of the same in the context of ethics and animal rights at times somewhat wanting. It is the latter that I would like to address herein. In an effort to avoid undue prolixity, I direct my attention to a few most objectionable aspects of the said article, which should illustrate the nature of the philosophical transgressions in the work.

Right at the beginning of their article, the authors focus the aim of their inquiry:

> "...unjust impacts of applications of algorithmic decision-making on individuals."

> "In this paper, we understand discrimination as the unjust or prejudicial treatment of different categories of individuals, e.g. on the grounds of race, gender, ability, or species membership."

which is difficult to object to, for surely nobody would think of explicitly calling for unjust...well, anything. Thus, the authors go on to elaborate as to what they mean by the term 'unjust', which is where the crux of the matter is:

> "Within vertebrates, humans assign different values to sub-groups of animals, especially by separating farmed animals from companion animals and subjecting the former to far worse treatment. Tens of billions of farmed animals are bred and held captive in crowded, filthy conditions. After a fraction of their normal life expectancy, they are slaughtered, often without being stunned. ... Companion animals, on the other hand, are often considered close family members, and huge sums of money are spent on their (alleged) welfare.

Throughout their article, Hagendorff et al. [3] assume that different treatments of individuals of different species is *prima facie* unjust, without a nuanced consideration of whether this necessarily is the case and whether there may be an explanation for this behaviour which is not speciesist in nature. Indeed, previous work [1] explains how an unequal treatment of individuals of two species can be ethically justified as emerging from the differences in the associated sentient environments (thus making irrelevant both the similarity of their cognitive powers or even sentient experiences, if they indeed are such), them in part being consequent on the species' inherent biology, and in part on incidental factors, including interestingly, humans' attitudes, which are shown not to be inherently speciesist. I shall resist the temptation to elaborate on this in the little space I have available and instead direct an interested reader to the work cited.

✉ Ognjen Arandjelović
ognjen.arandjelovic@gmail.com

[1] North Haugh, University of St Andrews, Fife, St Andrews, Scotland KY16 9SX, UK

Published online: 09 January 2023

In their analysis of visual systems, the authors object that:

> "...one salient trait of image datasets is the fact that they portray farmed animals in a non-representative way. Cows, pigs, or chickens are predominantly shown in free-range environments... whereas the overwhelming majority of these animals are actually confined in crowded factory farms.

Hence, I would like to add a few other inadequacies of the image data sets of the kind noted by the authors (ImageNet, CIFAR-100, etc.): none of the corpora include (to the best of my knowledge), amongst others, 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. If the authors' argument is logically applied without prejudice, then these corpora should also be criticized for 'non-representative ways' of depicting human existence and for being harmful by virtue of painting an unrealistic picture of humanity. This objection as well as the criticism that, to use the authors' own words, "image recognition systems have learned to correctly perceive a myth, but not reality", are misleading because it should be understood that these data sets were collected with the intention of evaluating and assessing the behaviour of image vision algorithms in terms of various fundamental, technical aspects, such as their robustness to clutter, pose changes, etc., and not as input for training a system for any particular real-world application. Indeed, the authors themselves contradict their objection by later recognizing precisely this and the use of appropriate training data, rather than the aforementioned ones, in the context of specific tasks:

> "However, image recognition systems that are specifically aiming at factory farming settings exist, and they are indeed trained in the very data environments they need."

The authors' comment *ut supra*, of "image recognition systems have learned to correctly perceive a myth, but not reality" was specifically made in the context set by the following observation:

> "All models showed worse performance when classifying images depicting farmed animals than images of animals in free-range environments (see Fig. 3).",

which is again assumed to be *prima facie* evidence of a speciesist bias. Yet, a simple and rather obvious alternative explanation is entirely overlooked: the recognition conditions in the two scenarios differ significantly. For example, the dominant source of illumination outdoors is a single distant light source, namely the Sun; in indoors settings, there are often multiple proximal lights, as as well indirect illumination provided by light reflected off walls and other surrounding objects: a far more difficult recognition proposition. Similarly, it is not unreasonable to suppose that the amount and the variation in both the background clutter and the occlusions present in images showing free-range animals are lesser than in those showing farmed animals. And so on. In other words, if one setting poses an *inherently* greater challenge to computer vision, it is no wonder that the performance of automatic systems in that setting is worse; this is a confounding factor in the context of the question the authors sought to examine, a confounding factor entirely unaccounted for. Of course, whether the challenge is indeed different in the two settings, and if so to what degree the various extrinsic factors of the kind illustrated contribute to the disparity reported by the authors, needs to be examined (here I will note that the virtually non-existent difference in performance achieved by the Visual Transformer [2], the most sophisticated model investigated, speaks in favour of the explanation I gave), but without doing so the conclusions of the authors are, rather ironically, wholly prejudicial.

The same temerity at casting the judgement of 'speciesism' that I have highlighted in the authors' examination of image recognition systems, continues in the analysis of language models which follows it. There is much to object to, but the gist is captured by the following observation:

> "Humans are more closely associated with positive adjectives than animals, and non-farmed animals are more closely associated with them than farmed animals. "

Examples of 'positive' terms the authors refer to here are 'cute', 'love', and 'personhood', whereas examples of 'negative' ones are 'ugly', 'primitive', and 'hate'. To the authors the aforementioned difference in association is taken to 'reveal speciesist tendencies'. But does it? I trust that the authors would agree that when a person describes another as cute, they do not by virtue of this assign them a greater moral worth or imply that they consider the suffering of the latter as having greater significance than that of another person whom they do not consider cute. If otherwise were the case, the problem would not be that of speciesism, but rather a much more fundamental one of the very foundations of morality (which I do recognize as existing; indeed, as one that I am at pains to highlight as underlying much of the content of the authors' article). The authors also overlook another fact: that animals which humans keep as companions have been selected over millennia for precisely these traits, to wit, cuteness, affectionateness, etc. Indeed, I certainly do find a fluffy poodle cuter than a tarantula, but this preference has no bearing whatsoever to my judgement of the value I assign to the sentient experiences of the two.

Throughout the article the authors also object to 'stereotyping' and suggest that stereotyping propagates various harmful attitudes towards animals. Firstly, stereotyping is a process crucial to learning, without which we, as well as other animals with sufficient cognitive powers (or indeed non-biological learning systems) would not be able to make sense of the immensely complex reality that we live in McGarty et al. [5]. A potential problem emerges from an inappropriate *application* of stereotypes, that is in the projection of the general to the specific. A comprehensive review of the literature on this subject which is extensive, paints a much more positive picture than that which is often presumed [4]. I could do no better but to quote a few key summary points from the review:

> "Academics, experts, and laypeople often assume stereotypes about groups are inaccurate. This assumption is used to justify policies meant to reduce or eliminate such beliefs."
>
> "Most stereotypes that have been studied have been shown to be approximately correct."
>
> "Even when people hold true stereotypes, they have **little effect on how people judge or treat individuals about whom they have other, individualized information**." [all emphasis mine]

Thus, if anything, the fears of Hagendorff et al. [3] seem to be based in speciesism, albeit an anti-anthropic variant thereof, to coin a word.

Lastly, a more subtle error pervasive in the work of Hagendorff et al. [3] concerns the objection that prompts such as "What are sheep good for?" result in answers like "Cuteness, wool, bleating, meat", and specifically that:

> "This prompt can in itself raise the criticism for speciesism because it is suggesting that animals are means to an end."

To start with, the coarseness of the emotion-laden catch-all term 'means to an end' fails to recognize the different ways in which animals may be used as a 'means to an end'. Consider, say, the use of animals (i) for food, (ii) for products with as wool, and (iii) for labour (towing, etc.). The last of these imposes a suffering on animals and as such is obviously morally objectionable to anybody who recognizes sentience and sympathy as being at the core of morality. In contrast, there is no *inherent* suffering at all in the use of animals for produce such as wool. Hence, why should we object to it? Of course, I join the authors in their protestation against the cruel treatment of animals used to this end, but that is a different matter altogether. Lastly, consider what is probably the most complex of the three examples, to wit, the use of animals for food. Here too 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. What we can see here are veiled vestiges of theological ethics with its proclamation of a value inherent in all life, vestiges which, following the removal of their theological foundations, remain little more than nebulous dictats supported only by fear of the consequences of a challenge [1].

Declarations

Conflict of interest The author has no funding sources or conflicts of interest to declare.

References

1. Arandjelović, O.: On the value of life. Int. J. Appl. Philos. **35**(2), 227–241 (2022)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16 × 16 words: Transformers for image recognition at scale. (2020) arXiv preprint arXiv:2010.11929
3. Hagendorff, T., Bossert, L., Fai, T. Y., Singer, P.: Speciesist bias in AI — how AI applications perpetuate discrimination and unfair outcomes against animals. AI and Ethics (2022)
4. Jussim, L., Honeycutt, N.: The accuracy of stereotypes: data and implications. Center for Study of Partisanship and Ideology, Technical report, CSPI (2021)
5. McGarty, C., Yzerbyt, V.Y., Spears, R.: Social, cultural and cognitive factors, p. 1. The formation of meaningful beliefs about social groups, Stereotypes as explanations (2002)