philosophies

MDPI

Article

AI Ethics and Value Alignment for Nonhuman Animals

Soenke Ziesche

AI Policy Labs, New Delhi 110001, India; soenke.ziesche@aipolicylabs.org

Abstract: This article is about a specific, but so far neglected peril of AI, which is that AI systems may become existential as well as causing suffering risks for nonhuman animals. The AI value alignment problem has now been acknowledged as critical for AI safety as well as very hard. However, currently it has only been attempted to align the values of AI systems with human values. It is argued here that this ought to be extended to the values of nonhuman animals since it would be speciesism not to do so. The article focuses on the two subproblems—value extraction and value aggregation—discusses challenges for the integration of values of nonhuman animals and explores approaches to how AI systems could address them.

Keywords: AI safety; AI ethics; AI value alignment problem; value extraction; value aggregation; nonhuman animals; speciesism

check for updates

Citation: Ziesche, S. AI Ethics and Value Alignment for Nonhuman Animals. *Philosophies* **2021**, *6*, 31. https://doi.org/10.3390/philosophies6020031

Academic Editor: Gordana Dodig-Crnkovic

Received: 23 March 2021
Accepted: 6 April 2021
Published: 13 April 2021

1. Introduction

As there has been significant progress of AI applications in a range of fields with some delay, the fields of AI ethics [1] and safety [2] have also gained traction. AI ethics and safety concern a variety of issues, of which the value alignment problem may be the most important—as well as, perhaps, the hardest [3,4]. Very briefly, it is about ensuring that AI systems, especially those not yet developed super intelligent Artificial General Intelligence systems, pursue goals and values which are aligned with human goals and values.

In this article the term "human" in the above definition is challenged since this implies that it would not matter if AI systems were not necessarily aligned with the interests of other beings. In particular, this article focuses on the interests of nonhuman animals. Thus, we argue for widening the above definition to the goals and values of humans as well as of nonhuman animals.

In the history of humankind many ethical views have changed and developed. Bostrom and Yudkowsky pointed out that AI ethics should not be stable, but be open for changes if humans recognize previous ethical mistakes [5]. As will be outlined, the treatment of nonhuman animals is an example of an ethical issue which has been changed over time or is in the process of changing. Therefore, we aim to incorporate nonhuman animals into the ongoing research of AI ethics.

Apart from a very small number of attempts, deliberations of ethical obligations of humans towards nonhuman animals gained momentum only in the late 20th century [6], for example through the approach of contractualism [7]. The main criterion was the recognition that (many species of) nonhuman animals are sentient, thus able to experience suffering [8]. This also led to the propagation of the term "speciesism", which Horta defines as "unjustified disadvantageous consideration or treatment of those who are not classified as belonging to a certain species" [9] (p. 1).

The progress on the moral side has also begun to be reflected in regulations. There are nowadays in a number of countries whose national laws recognize nonhuman animal sentience and suffering (https://www.globalanimallaw.org/database/national/index.html, accessed on 11 April 2021), while a "Universal Declaration on Animal Welfare" is still at a proposal stage (https://www.globalanimallaw.org/database/universal.html, accessed on 11 April 2021) and global agreements such as the Convention on Biological Diversity

(https://www.cbd.int/, accessed on 11 April 2021) from 1992 and the United Nations Sustainable Development Goals from 2015, especially goals 14 and 15 [10], rather focus on the protection of species (from becoming extinct) than on their explicit welfare.

We argue that AI systems would be better custodians for nonhuman animals than humans for the following reasons:

Humans are biased and inconsistent when it comes to the treatment of nonhuman animals since many humans have an interest in their meat, eggs, milk, fur, leather, wool, etc. as well in the habitats of nonhuman animals to use them for their own purposes.

Humans have incomplete knowledge about how to ensure animal welfare, which includes the largely neglected suffering of wild animals [11], while AI systems may find solutions in this regard.

Up to now nonhuman animals have hardly been taken into consideration for the AI value alignment problem. For example, the so-called Asilomar Principles (https://futureoflife.org/ai-principles/, accessed on 11 April 2021) towards a beneficial AI by leading AI researchers focus only on humans:

"(10) Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.

(11) Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity."

On the platform "Linking Artificial Intelligence Principles" [12] reference to (nonhuman) animals can only be found in the "Draft Ethics Guidelines for Trustworthy AI" by the European Commission's High-Level Expert Group on Artificial Intelligence (https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=57112, accessed on 11 April 2021). However, in the final document [13] the text has been removed, which stated in the draft: "Avoiding harm may also be viewed in terms of harm to the environment and animals, thus the development of environmentally friendly AI may be considered part of the principle of avoiding harm".

For technological design in general an approach called Value Sensitive Design has been introduced, which incorporates human values at every stage [14] and which has been recently extended to the values of nonhuman animals [15]. Baum borrows the term "standing" from legal contexts to pose the question "Who or what is included in the group to have its values factored into the AI?" [16] (p. 4). He then elaborates on challenges for certain groups, such as antisocial psychopaths, but also children as well as nonhuman animals. Additionally, Gabriel mentions in his elaboration on AI alignment nonhuman animals and sentient beings in general initially, but not anymore when he explores the details of value extraction and aggregation [17]. Sarma and colleagues argue for an interdisciplinary approach for the AI value alignment problem, involving neuroscience in particular, which may lead to insights of what they call "mammalian value systems" [18,19]. Ziesche and Yampolskiy raise the issues of speciesism and animal welfare, but their main focus is on sentient digital minds [20].

Yudkowsky used the following quote to illustrate that if the value alignment problem is not taken seriously there is no reason to believe that sufficiently powerful AI systems may not harm humans: "The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else." [4] (p. 27). Yet, this quote can be applied also to nonhuman animals for AI systems, which have not been aligned with the preferences of nonhuman animals. In other words, such neglect would lead to existential risks [21] and suffering risks [22] for nonhuman animals. Since we would consider this as a serious peril of AI the purpose of this article is to fill this lacuna and not only to confirm the standing of nonhuman animals [16], but also to enhance the limited groundwork that has been done so far. Whereas previous research has explored the ethical obligations towards nonhuman animals, AI value alignment for humans as well as, at an early stage, technological design for nonhuman animals, this article is comparatively unique given that

it brings together various existing approaches to look specifically at two subproblems of the AI value alignment problem related to nonhuman animals.

The complex AI value alignment problem can be broken down into at least three subproblems, of which only the first two are discussed here. First, it has to be attempted to precisely extract the values of all concerned beings; second, it has to be attempted to aggregate these values in a consistent manner; third, it has to be attempted to instill these values into AI systems. The first two subproblems have in common that they are normative, while the third one is technical [17].

To address these topics this article has three further sections. Section 2 explores the nature of the value extraction and Section 3 the nature of the value aggregation. These sections are divided into five subsections each: introduction, challenges, non-AI activities on value extraction/aggregation for nonhuman animals, potential AI approaches and summary. In the final section, Section 4, an overall summary and an outlook at future work is provided.

2. Value Extraction

2.1. Introduction

For the value extraction subproblem, which can be considered as a modern version of centuries-old ethical questions in philosophy towards right and wrong behavior, first the terminology has to be clarified [3]. Gabriel introduced the following six options to describe what we may want an AI system or agent to do [17] (pp. 5–9).

i. Instructions: the agent does what I instruct it to do.

ii. Expressed intentions: the agent does what I intend it to do.

iii. Revealed preferences: the agent does what my behavior reveals I prefer.

iv. Informed preferences or desires: the agent does what I would want it to do if I were rational and informed.

v. Interest or wellbeing: the agent does what is in my interest, or what is best for me, objectively speaking.

vi. Values: the agent does what it morally ought to do, as defined by the individual or society."

These wordings also illustrate well some challenges of incorporating the values of nonhuman animals into the AI value alignment problem; hence, we will revert to them below. The first option to explicitly instruct the AI has been, after initial attempts, by now widely dismissed. It would be too complex, may lead to perverse instantiations and may not be flexible for changes. For the second option the potential issue is that the creator may have unethical, irrational or misinformed intentions. Similarly for the third option, the preferences may actually be harmful for the creator as well as for other people and may have been developed based on incomplete information. The fourth option, which resembles Yudkowsky's approach of "Coherent Extrapolated Volition" (CEV) [23], and the fifth option address the shortcomings of the second and third option, but also have challenges. Therefore, Gabriel favors the sixth option that the AI does what it morally ought to do, but explains that the challenges remain to specify the values and to aggregate them [17].

2.2. Challenges

When incorporating nonhuman animals to the value alignment problem the following additional challenges regarding the subproblem of value extraction emerge:

To actually identify the values and interests of nonhuman animals, taking into account that humans do not represent nonhuman animals because the values and preferences of humans and nonhuman animals are not consistent because, for example, many humans eat certain nonhuman animals and destroy the habitats of nonhuman animals.

To explore a large variety of species of nonhuman animals, thus likely a large variety of values and interests and to potentially draw a line as to which nonhuman animals are not to be considered for AI value alignment.

To focus on endangered species, assuming that self-preservation is a preference of all nonhuman animals.

2.3. Non-AI Activities on Value Extraction for NonHuman Animals

As Baum pointed out "human AI designers do not know what it is like to be a cow or a frog, so they must make sweeping assumptions in programming an AI to infer cow or frog ethics from any given measurement technique" [16] (p. 11). In this regard we propose animal welfare science as a suitable field of study to analyze interests of nonhuman animals, which has been also fairly recently introduced [24,25].

For humans an essential method to inform about their wellbeing is self-reporting through direct communication, which is challenging to be applied to nonhuman animals, although attempts are being made [26]. Instead, animal welfare science mostly relies on quantitative information through functional (physiological) as well as through behavioral indicators. Sensing technologies offer a variety of options to measure a range of indicators, either through wearables or remotely [27]. In this regard the challenging issue of consent of nonhuman animals has to be addressed [28]. Passive integrated transponder devices (PIT tags), which are microchips injected into nonhuman animals, provide both types of data—behavioral, such as movements, as well as physiological, such as the body temperature [29].

Examples for functional indicators are health vitals, but also emotions such as pain [30]. The most accurate thing to do would be to measure the emotions through corresponding brain processes, but science is not advanced enough yet [31]. The alternatives are indirect measurements, such as the hypothalamic–pituitary–adrenal function to study stress in nonhuman animals [32]. Overall, the studies of functional indicators as well as wearable technologies are more sophisticated and more often applied to domestic and farm animals and for practical reasons less so for wild animals, while for this article this distinction is of no relevance.

Behavioral indicators are concerned with movements and activities. M. S. Dawkins has defined what nonhuman animals strive for as positive reinforcers and what they try to avoid as negative reinforcers [33]. For some nonhuman animals the analysis of facial expressions supports the assessment of pain [34] as well as of positive emotions [35]. For wild animals, behavioral indicators measured through remote sensing systems are an appropriate means as well, as often the only feasible way to assess their welfare. Examples are camera traps on the ground [36] or sensing systems located on aircrafts or satellites [37]. PIT tags are also used for wild animals, although the required tentative capture and the tagging can be stressful for them [29].

Similar to the positive and the negative branch of the ethical theory of utilitarianism it can be distinguished whether to focus on the promotion or maximizing of the wellbeing of nonhuman animals or the reduction or minimizing of suffering of nonhuman animals. There are more efforts made in terms of research [33] as well as of advocacy [38] to address the pain, stress and distress of nonhuman animals; while within these efforts the focus is on farm animals, the immense suffering of wild animals also has to be highlighted [11], which is, unlike the suffering of farm animals, mostly not caused by humans. Yet, in addition to attempts towards the avoidance of suffering, research is also conducted on the pleasure of nonhuman animals and its moral significance [39,40], and evaluation methods [41] and indicators [42] in this regard are being developed.

Several categorizations of criteria for animal wellbeing have been developed [43], often again with a focus on farm animals [44]. One example are the five domains proposed by Mellor and Reid [45]: 1. Thirst/hunger/malnutrition; 2. Environmental challenge; 3. Disease/injury/functional impairment; 4. Behavioral/interactive restriction; 5. Anxiety/fear/pain/distress.

Regarding the second challenge of the high number of species of nonhuman animals, a threshold could be to consider only conscious or sentient nonhuman animals. However, this is a complex topic [46], which is beyond the scope of this article. For example, the

possibility of insect suffering has been looked at, which would be, due to the high numbers of insects, a significant portion of the overall suffering on earth [47]. As an approach to study the welfare of a very high number of species, it has been suggested to select and examine so-called indicator species and then generalize from these species to other species [43].

The third challenge carries a time component. If we assume that self-preservation is a preference of all species of nonhuman animals, regardless of whether they are conscious or not, then the conversation of endangered species has to be a priority. While extinctions of species are normal and have occurred regularly through the evolutionary process, current extinction rates are up to 1000 times higher than normal extinction rates due to the impact of humans [48]. The reasons for this high rate are, for example, overhunting and the destruction of habitats. The International Union for Conservation of Nature maintains the "Red List of Threatened Species" (https://www.iucnredlist.org/, accessed on 11 April 2021). There are international agreements on conversation, such as the abovementioned Convention on Biological Diversity, and also various NGOs and private initiatives are committed to conservation.

2.4. Potential AI Approaches

We introduce approaches for how value extraction for the AI value alignment problem could incorporate the interests of nonhuman animals taking into account the three challenges introduced above.

As outlined, research in animal welfare science provides a foundation to evaluate the wellbeing of nonhuman animals, but it has neither been linked to the AI value alignment problem nor have AI methods thoroughly been applied to animal welfare science.

Therefore, our main proposal is to provide AI systems access with to all types of data of nonhuman animals according to animal welfare methods and let the AI systems analyze the data. It is critical that the AI systems can access the raw data as opposed to human conclusions based on them since humans tend to be self-serving and biased, e.g., due to their interest in farm animals or in wildlife habitats, and less smart than the AI, and thus are likely to be only imperfectly informed about animal welfare. In contrast, AI systems may discover hidden or unknown patterns related to the wellbeing of nonhuman animals.

As described, a wealth of functional and behavioral data of nonhuman animals is already available or can be collected, including PIT data and satellite/camera footage of wild animals (e.g., https://www.wildlifeinsights.org, accessed on 11 April 2021). The analysis capabilities of such sensory inputs using deep reinforcement learning [49] or inverse reinforcement learning [50] have advanced significantly and may lead AI systems to insights about the wellbeing of nonhuman animals, e.g., elephants [51], as has been suggested already related to the wellbeing of humans [52].

Moreover, AI systems may be capable of harnessing those methods of animal welfare science which are not yet advanced, such as direct communication with nonhuman animals, which would enable self-reporting of values and interests [53], or the measurement of wellbeing through correlated brain processes [19]. The measurement of brain processes would provide for optimal functional indicators, while the above ones are mostly proxy indicators.

For the second challenge regarding the large variety of species of nonhuman animals, AI systems may be able, again based on a large amount of data and perhaps direct communication, to ascertain which species are sentient, and thus morally matter. Therefore, the interests of all these species should be considered. However, as mentioned, this topic goes beyond the scope of this article, also given that the impartial ascertainment of sentience of beings would likely require rather advanced AI.

When it comes to endangered species the situation is different in the sense that the value, which is self-preservation, is clear and that a value-aligned AI system would focus on conservation without losing time. There are AI approaches towards conservation [54],

including the prevention of poaching [55], based on the analysis of behavioral data, as introduced above.

2.5. Summary

The introduced approaches can be summarized by recapping three of Gabriel's wordings [17]:

i. Expressed intentions: the agent does what I intend it to do.

This option, if projected on nonhuman animals, would mean that the AI does what nonhuman animals intend to do. However, this requires self-reporting of their interests, which is challenging, unless AI systems develop ways to communicate with nonhuman animals.

ii. Revealed preferences: the agent does what my behavior reveals I prefer.

For this option, research on behavioral indicators of animal welfare has been introduced.

iii. Interest or wellbeing: the agent does what is in my interest, or what is best for me, objectively speaking.

For this option, research on functional indicators of animal welfare has been introduced, ideally culminating in the identification of correlated brain processes for wellbeing.

For AI activities towards endangered species the setting differs since we assume the interest of self-preservation of all species as given, thus directly instructed to AI systems.

In summary, the introduced methods of animal welfare science reveal mostly interests of individual nonhuman animals as opposed to overall values of a whole species, let alone overall values of all nonhuman animals. Many preferences of nonhuman animals tend to be selfish, while only humans, despite the mentioned conflicting interests, may have concepts of overall welfare for nonhuman animals. However, such overall ethical considerations of nonhuman animals are critical for the subsequent subproblem of value aggregation.

3. Value Aggregation

3.1. Introduction

If all the challenges of the value extraction subproblem could be overcome, the ideal scenario would be that all the extracted values are compatible, in which case there would be no value aggregation subproblem. Yet, in our diverse world with a variety of human values this is not the case, and even less so if nonhuman animals are incorporated as shown below. Additionally, this problem has been a long-standing philosophical debate, mostly for humans only [3]. Due to the immense challenges, even personal universes have been envisaged as a solution [56]. Two main categories of approaches can be identified, top-down and bottom-up [57].

A top-down approach implies that, based on the outcome of an ideally fair value extraction, a decision is made about what ethical views the AI system should use. For example, despite our pluralistic world, humans can agree on universal human rights [58].

Bottom-up approaches are also referred to as social choice and aim for AI systems which integrate individual views. These approaches carry the advantages of being democratic as opposed to autocratic structures, yet also a range of challenges regarding practicability and fairness [16,17,57].

A supplemental approach is CEV, which was mentioned above for individual value extraction. Applied to value aggregation this means to ideally harness the intelligence of the AI system to have such ethical views we would want "if we knew more, thought faster, were more the people we wished we were, and had grown up farther together" [23] (p. 6) [3].

3.2. Challenges

When incorporating nonhuman animals to the value alignment problem the following additional challenges regarding the subproblem of value aggregation emerge:

To address the fact that values and interests of humans and nonhuman animals are conflicting, e.g., that many humans eat certain nonhuman animals and destroy habitats of nonhuman animals (see also the third challenge above) or, vice versa, that certain nonhuman animals kill humans.

To address the fact that certain short-term preferences of nonhuman animals are not necessarily good for their long-term health, e.g., that nonhuman animals may eat unhealthy food or lick and scratch wounds.

To address the fact that values and interests between species of nonhuman animals are not only different (see above), but also conflicting, e.g., that predators kill and consume prey.

3.3. Non-AI Activities on Value Aggregation for NonHuman Animals

Regarding the conflicting values and interests of humans and nonhuman animals, the practice of vegetarianism can be traced back to ancient times [59], yet factory farming is still prevalent and, related to it, suffering is caused to nonhuman animals [38].

Additionally, humans continue to cause the loss of biodiversity and extinction of a large number of species [60]. However, thousands of humans are also killed every year by nonhuman animals and since ancient times humans have tried to protect themselves in this regard, while any type of reconciliation between the species through value aggregation seems unthinkable.

The second challenge about bad choices related to probabilistic outcomes can be observed for humans [61] as well as for nonhuman animals, e.g., pigeons [62]. This issue is especially critical when it comes to (long-term) health, which should be considered as a fundamental element of animal welfare [33]. Therefore, some behavioral indicators introduced above to extract values towards animal welfare may not be sufficient for optimal value aggregation if humans and AI systems have the knowledge that certain behavior of nonhuman animals may be pleasurable in the short run, but not healthy in the long run, thus are actually deceptive indicators for animal welfare.

The baseline for the third challenge is the above insight that preferences of nonhuman animals tend to be about the welfare of themselves and their kin, yet oblivious to universal moral values. This is linked to the issue is that "evolution optimizes for reproductive success rather than individual wellbeing", which leads to enormous suffering of nonhuman animals [63]. In the on-going debate about whether humans may attempt to intervene in nature to reduce suffering of wild animals there are proponents [63,64] and opponents [65]. While the opponents, for example, warn of unintended ecological consequences, Tomasik states that "cancer, malaria, sexual violence, and depression are all 'natural' outcomes of evolution's optimization process, yet we rightly consider them evils to be resisted; we should encourage people to realize that the cruelties that nature inflicts on its nonhuman inhabitants are just as ethically intolerable—indeed more so, since the number of organisms affected in the latter case is orders of magnitude higher" [63]. For those who support interventions to reduce the suffering of wild animals, the question arises of how these interventions could be implemented without harm to the ecosystem.

While activities on a smaller scale, such as rescuing trapped and injured animals, healing sick animals or caring for orphaned animals are on-going and not debated, Pearce's "Abolitionist Project" takes a more comprehensive approach, which calls for the use of technology, such as genetic engineering, to abolish the suffering of nonhuman animals (as well as of humans) and which, due to its technical challenges and due to its controversy, has not commenced implementation [66,67].

3.4. Potential AI Approaches

We introduce approaches to how value aggregation for the AI value alignment problem could incorporate the interests of nonhuman animals while taking into account the three challenges introduced above.

While acknowledging the difficulties of the top-down approach, the fairly successful adoption of universal human rights within a diverse world of human values was introduced

above [58], and a human rights-congruent AI could be an option for AI value alignment for humans. In a similar manner, an AI congruent with human as well as with nonhuman animal rights could be considered as desirable, while, as mentioned, the third subproblem of the AI value alignment problem (i.e., how to instill these values into AI systems) is beyond the scope of this article (this would, given that it is solvable, ensure that AI systems respect the rights of both humans and nonhuman animals, which is similar to the status quo of the current world).

However, the question arises of whether we should not aim for more by attempting to let AI systems address the second and third challenge of the value aggregation problem above for nonhuman animals too, which are, in short, that nonhuman animals make choices that are not in their long-term interests and that many nonhuman animals cause other nonhuman animals to suffer. In other words, similarly to the approaches of AI value alignment for humans only, we would not merely aim to preserve the (not ideal) status quo of nonhuman animals, but to improve the current condition by looking at bottom-up as well as CEV approaches and by benefitting from the much larger potentials of the AI systems.

The CEV approach appears suitable to tackle bad the choices of nonhuman animals. It has been so far mostly explored for humans, but Bostrom posed the question of whether the volition of (higher) animals should be considered [3] (p. 216). Therefore, if an AI system, owing to its intelligence, can figure out long-term values of nonhuman animals, then it could override during value aggregation the less smart preferences of nonhuman animals, which it has extracted through the analysis of the behavioral data of nonhuman animals as proposed above.

The issue of (wild) animal suffering is linked to the above finding that the individual preferences of nonhuman animals are rather selfish, while overall animal utility approaches are desirable, which can probably not be conceived by nonhuman animals, but by humans and perhaps even better by AI systems. To tackle this, AI systems may support required technologies as conceptualized by Pearce [66,67]. This entails two issues—that AI systems are not only aware of the conflicting interests of the concerned wild animals, but are also solution-oriented and willing to apply their much greater intelligence in this regard. If this was successful, e.g., if predation could be prevented, this would also contribute to the first challenge above regarding the conflicting interests of humans and nonhuman animals in such a way that humans would not be killed anymore by nonhuman animals, thus contributing to the safety of humans. However, the suffering of nonhuman animals is not only caused by predators, but by various other causes triggering pain [11]. This could potentially be addressed by AI systems developing specialized drugs, a field that is in early, but promising stages for humans [68,69].

3.5. Summary

Firstly, the introduced approaches can be again summarized by recapping two further wordings of Gabriel [17]:

iv. Informed preferences or desires: the agent does what I would want it to do if I were rational and informed.

This option is suitable for the second challenge above and we introduced a proposal that AI systems may apply CEV.

v. Values: the agent does what it morally ought to do, as defined by the individual or society.

This option is suitable for the sixth challenge above and we introduced proposals how AI systems may intervene in wild animal suffering.

In summary, it is a complex field and goes beyond the scope of this article how to aggregate the values of humans and all nonhuman animals and to establish what are actually desirable conditions in the long run and whether interventions would cause more good than harm. However, an AI system may be a valuable advisor, provided it is a well aligned, e.g., by applying CEV. An outcome that has to be avoided is that any AI system

addresses the value aggregation for nonhuman animals through perverse instantiation [3] in a way that (painless) annihilation of all nonhuman animals would also relieve them from any suffering. This could be done through a critical weighting of the value of self-preservation.

4. Conclusions

This article aimed to create awareness of a specific peril of AI, which is that without appropriate actions AI systems may become existential as well as causing suffering risks for nonhuman animals. In this regard the contributions of this article are as follows:

We discussed why the AI value alignment should also be concerned with the values, preferences and interests of nonhuman animals, in addition to the values, preferences and interests of humans.

We outlined the challenges for two subproblems—value extraction and value aggregation—if these are extended to nonhuman animals, as well as potential AI approaches to address these challenges, e.g., by applying methods of animal welfare science.

Although the above points add another layer of complication to the already very hard AI value alignment problem, we believe it is ethically indispensable to include them since dismissing them would be speciesism and could increase the suffering of nonhuman animals or could lead to their extermination.

This article does not intend to provide solutions for longstanding problems such as suffering of wild animals or answers to the questions of how to minimize overall net suffering and what ideal states of wellbeing for nonhuman animals are, but the article aims to look at methods so that AI systems become aware of as well as supportive of these issues.

As for future work, it may have to be considered that there are further morally relevant entities, which ought to be incorporated to AI value alignment too. Examples are the environment [17], which is also linked to the wellbeing of humans as well as nonhuman animals, or potentially other minds apart from humans and nonhuman animals [20].

Funding: This research received no external funding.

Acknowledgments: The author would like to thank two anonymous reviewers for helpful and inspiring feedback.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Müller, V.C.; Ethics of Artificial Intelligence and Robotics. In The Stanford Encyclopedia of Philosophy (Winter 2020 Edition); Zalta, E.N., Ed.; 2020. Available online: https://plato.stanford.edu/archives/win2020/entries/ethics-ai. (accessed on 11 April 2021).
2. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. *arXiv* **2016**, arXiv:1606.06565.
3. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
4. Yudkowsky, E. Artificial intelligence as a positive and negative factor in global risk. In *Global Catastrophic Risks*; Bostrom, N., Ćirković, M.M., Eds.; Oxford University Press: New York, NY, USA, 2008; pp. 308–345. Available online: https://intelligence.org/files/AIPosNegFactor.pdf (accessed on 10 March 2021).
5. Bostrom, N.; Yudkowsky, E. The ethics of artificial intelligence. In *The Cambridge Handbook of Artificial Intelligence*; Frankish, K., Ramsey, W.M., Eds.; Cambridge University Press: Cambridge, UK, 2014; pp. 316–334. Available online: http://faculty.smcm.edu/acjamieson/s13/artificialintelligence.pdf (accessed on 10 March 2021).
6. Rollin, B.E. The regulation of animal research and the emergence of animal ethics: A conceptual history. *Theor. Med. Bioeth.* **2006**, *27*, 285–304. Available online: https://org.uib.no/dyreavd/handouts/Rollin__B._2006._Animal_Research_Regulation_in_Theoret._Medicin_....PDF (accessed on 10 March 2021). [CrossRef] [PubMed]
7. Scanlon, T. *What We Owe to Each Other*; Belknap Press: Cambridge, MA, USA, 2000.
8. Singer, P. *Animal Liberation: A New Ethics for Our Treatment of Animals*; HarperCollins: New York, NY, USA, 1975.
9. Horta, O. What is speciesism? *J. Agric. Environ. Ethics* **2010**, *23*, 243–266.
10. United Nations General Assembly. Transforming Our World: The 2030 Agenda for Sustainable Development. *Resolution A/RES/70/1*. 2015. Available online: https://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_RES_70_1_E.pdf (accessed on 10 March 2021).

11. Tomasik, B. The importance of wild-animal suffering. *Relat. Beyond Anthr.* **2015**, *3*, 133. Available online: https://www.ledonline.it/index.php/Relations/article/download/880/717 (accessed on 10 March 2021). [CrossRef]
12. Zeng, Y.; Lu, E.; Huangfu, C. Linking artificial intelligence principles. *arXiv* **2018**, arXiv:1812.04814.
13. High-Level Expert Group on AI. Ethic Guidelines for Trustworthy AI. 2019. Available online: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 (accessed on 10 March 2021).
14. Friedman, B.; Hendry, D.G. *Value Sensitive Design: Shaping Technology with Moral Imagination*; MIT Press: Cambridge, MA, USA, 2019.
15. Umbrello, S. The Ecological Turn in Design: Adopting a Posthumanist Ethics to Inform Value Sensitive Design. *Philosophies* **2021**, *6*, 29. Available online: https://www.mdpi.com/2409-9287/6/2/29/htm (accessed on 5 April 2021). [CrossRef]
16. Baum, S.D. Social choice ethics in artificial intelligence. *AI Soc.* **2020**, *35*, 165–176. Available online: https://sethbaum.com/ac/2020_SocialChoice.pdf (accessed on 10 March 2021).
17. Gabriel, I. Artificial intelligence, values, and alignment. *Minds Mach.* **2020**, *30*, 411–437. Available online: https://link.springer.com/content/pdf/10.1007/s11023-020-09539-2.pdf (accessed on 10 March 2021).
18. Sarma, G.; Hay, N. Mammalian value systems. *Informatica* **2017**, *41*. Available online: https://arxiv.org/pdf/1607.08289.pdf (accessed on 10 March 2021). [CrossRef]
19. Sarma, G.P.; Safron, A.; Hay, N.J. Integrative biological simulation, neuropsychology, and AI safety. *arXiv* **2018**, arXiv:1811.03493. [CrossRef]
20. Ziesche, S.; Yampolskiy, R.V. Towards AI Welfare Science and Policies. *Spec. Issue Artif. Superintelligence Coord. Strategy Big Data Cogn. Comput.* **2018**, *3*, 2. Available online: https://www.mdpi.com/2504-2289/3/1/2/htm (accessed on 10 March 2021).
21. Bostrom, N. Existential risks: Analyzing human extinction scenarios and related hazards. *J. Evol. Technol.* **2002**, *9*. Available online: https://nickbostrom.com/existential/risks.html (accessed on 10 March 2021).
22. Althaus, D.; Gloor, L. Reducing Risks of Astronomical Suffering: A Neglected Priority. 2016. Available online: https://foundational-research.org/reducing-risksofastronomical-suffering-a-neglected-priority/ (accessed on 10 March 2021).
23. Yudkowsky, E. *Coherent Extrapolated Volition*; The Singularity Institute: San Francisco, CA, USA, 2004. Available online: https://intelligence.org/files/CEV.pdf (accessed on 10 March 2021).
24. Broom, D.M. Animal welfare: Concepts and measurement. *J. Anim. Sci.* **1991**, *69*, 4167–4175. [CrossRef]
25. Broom, D.M. A history of animal welfare science. *Acta Biotheor.* **2011**, *59*, 121–137. [CrossRef]
26. Carey, M.P.; Fry, J.P. Evaluation of animal welfare by the self-expression of an anxiety state. *Lab. Anim.* **1995**, *29*, 370–379. [CrossRef]
27. Jukan, A.; Masip-Bruin, X.; Amla, N. Smart computing and sensing technologies for animal welfare: A systematic review. *ACM Comput. Surv. CSUR* **2017**, *50*, 1–27. Available online: https://dl.acm.org/doi/pdf/10.1145/3041960 (accessed on 10 March 2021). [CrossRef]
28. Mancini, C. Towards an animal-centred ethics for Animal–Computer Interaction. *Int. J. Hum. Comput. Stud.* **2017**, *98*, 221–233. Available online: http://oro.open.ac.uk/46164/1/Mancini_1-s2.0-S1071581916300180-main.pdf (accessed on 10 March 2021). [CrossRef]
29. Gibbons, W.J.; Andrews, K.M. PIT tagging: Simple technology at its best. *Bioscience* **2004**, *54*, 447–454. Available online: https://academic.oup.com/bioscience/article/54/5/447/417483 (accessed on 10 March 2021). [CrossRef]
30. Neethirajan, S.; Reimert, I.; Kemp, B. Measuring Farm Animal Emotions—Sensor-Based Approaches. *Sensors* **2021**, *21*, 553. Available online: https://www.mdpi.com/1424-8220/21/2/553/htm (accessed on 10 March 2021). [CrossRef]
31. Paul, E.S.; Harding, E.J.; Mendl, M. Measuring emotional processes in animals: The utility of a cognitive approach. *Neurosci. Biobehav. Rev.* **2005**, *29*, 469–491. [CrossRef]
32. Mormède, P.; Andanson, S.; Aupérin, B.; Beerda, B.; Guémené, D.; Malmkvist, J.; Manteca, X.; Manteuffel, G.; Prunet, P.; van Reenen, C.G.; et al. Exploration of the hypothalamic–pituitary–adrenal function as a tool to evaluate animal welfare. *Physiol. Behav.* **2007**, *92*, 317–339. [CrossRef]
33. Dawkins, M.S. The science of animal suffering. *Ethology* **2008**, *114*, 937–945. Available online: http://users.ox.ac.uk/~{}snikwad/resources/eth1557.pdf (accessed on 10 March 2021). [CrossRef]
34. McLennan, K.M.; Miller, A.L.; Dalla Costa, E.; Stucke, D.; Corke, M.J.; Broom, D.M.; Leach, M.C. Conceptual and methodological issues relating to pain assessment in mammals: The development and utilisation of pain facial expression scales. *Appl. Anim. Behav. Sci.* **2019**, *217*, 1–15. Available online: https://chesterrep.openrepository.com/bitstream/handle/10034/622559/ChesterREP_KMcLennan_Review.pdf?sequence=4&isAllowed=n (accessed on 10 March 2021). [CrossRef]
35. Lansade, L.; Nowak, R.; Lainé, A.L.; Leterrier, C.; Bonneau, C.; Parias, C.; Bertin, A. Facial expression and oxytocin as possible markers of positive emotions in horses. *Sci. Rep.* **2018**, *8*, 14680. Available online: https://www.nature.com/articles/s41598-018-32993-z (accessed on 10 March 2021). [CrossRef]
36. Ahumada, J.A.; Fegraus, E.; Birch, T.; Flores, N.; Kays, R.; O'Brien, T.G.; Dancer, A. Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet. *Environ. Conserv.* **2020**, *47*, 1–6. Available online: https://www.cambridge.org/core/journals/environmental-conservation/article/wildlife-insights-a-platform-to-maximize-the-potential-of-camera-trap-and-other-passive-sensor-wildlife-data-for-the-planet/98295387F86A977F2ECD96CCC5705CCC (accessed on 10 March 2021). [CrossRef]

37. Wang, D.; Shao, Q.; Yue, H. Surveying wild animals from satellites, manned aircraft and unmanned aerial systems (UASs): A review. *Remote Sens.* **2019**, *11*, 1308. Available online: https://www.mdpi.com/2072-4292/11/11/1308/pdf (accessed on 10 March 2021). [CrossRef]
38. Harari, Y.N. Industrial Farming is One of the Worst Crimes in History. *The Guardian*. 25 September 2015. Available online: https://www.theguardian.com/books/2015/sep/25/industrial-farming-one-worst-crimes-history-ethical-question (accessed on 10 March 2021).
39. Balcombe, J. Animal pleasure and its moral significance. *Appl. Anim. Behav. Sci.* **2009**, *118*, 208–216. [CrossRef]
40. Balcombe, J. *Pleasurable Kingdom: Animals and the Nature of Feeling Good*; St. Martin's Press: New York, NY, USA, 2006.
41. Yeates, J.W.; Main, D.C. Assessment of positive welfare: A review. *Vet. J.* **2008**, *175*, 293–300. [CrossRef]
42. Wathes, C. Lives worth living? *Vet. Rec.* **2010**, *166*, 468–469. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1005.6278&rep=rep1&type=pdf (accessed on 10 March 2021). [CrossRef]
43. Brennan, O. "Fit and Happy": How Do We Measure Wild-Animal Suffering. 2017. Available online: https://was-research.org/paper/fit-happy-measure-wild-animal-suffering/ (accessed on 10 March 2021).
44. Botreau, R.; Veissier, I.; Butterworth, A.; Bracke, M.B.; Keeling, L.J. Definition of criteria for overall assessment of animal welfare. *Animal Welfare* **2007**, *16*, 225. Available online: https://www.researchgate.net/profile/Isabelle_Veissier/publication/40105884_Definition_of_criteria_for_overall_assessment_of_animal_welfare/links/0912f5086cafe81def000000.pdf (accessed on 10 March 2021).
45. Mellor, D.J.; Reid, C.S.W. Concepts of Animal Wellbeing and Predicting the Impact of Procedures on Experimental Animals. 1994. Available online: https://www.wellbeingintlstudiesrepository.org/cgi/viewcontent.cgi?article=1006&context=exprawel (accessed on 10 March 2021).
46. Allen, C.; Trestman, M. Animal consciousness. In *The Stanford Encyclopedia of Philosophy (Winter 2020 Edition)*; Zalta, E.N., Ed.; 2020. Available online: https://plato.stanford.edu/archives/win2020/entries/consciousness-animal/ (accessed on 11 April 2021).
47. Tomasik, B. The Importance of Insect Suffering. 2016. Available online: https://reducing-suffering.org/the-importance-of-insect-suffering/ (accessed on 10 March 2021).
48. De Vos, J.M.; Joppa, L.N.; Gittleman, J.L.; Stephens, P.R.; Pimm, S.L. Estimating the normal background rate of species extinction. *Conserv. Biol.* **2015**, *29*, 452–462. Available online: https://www.zora.uzh.ch/id/eprint/98443/1/Conservation_Biology_2014_early-view.pdf (accessed on 10 March 2021). [CrossRef]
49. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv* **2013**, arXiv:1312.5602. Available online: https://arxiv.org/pdf/1312.5602.pdf (accessed on 10 March 2021).
50. Ng, A.Y.; Russell, S. Algorithms for inverse reinforcement learning. In Proceedings of the Seventeenth International Conference on Machine Learning, Stanford, CA, USA; 2000. Available online: https://ai.stanford.edu/~{}ang/papers/icml00-irl.pdf (accessed on 12 April 2021).
51. Duporge, I.; Isupova, O.; Reece, S.; Macdonald, D.W.; Wang, T. Using very-high-resolution satellite imagery and deep learning to detect and count African elephants in heterogeneous landscapes. *Remote Sens. Ecol. Conserv.* **2020**. Available online: https://zslpublications.onlinelibrary.wiley.com/doi/full/10.1002/rse2.195 (accessed on 10 March 2021). [CrossRef]
52. Russell, S. Should we fear supersmart robots? *Sci. Am.* **2016**, *314*, 58–59. Available online: http://aima.cs.berkeley.edu/~{}russell/papers/sciam16-supersmart.pdf (accessed on 10 March 2021). [CrossRef] [PubMed]
53. Bjorck, J.; Rappazzo, B.H.; Chen, D.; Bernstein, R.; Wrege, P.H.; Gomes, C.P. Automatic detection and compression for passive acoustic monitoring of the African forest elephant. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 476–484.
54. Lamba, A.; Cassey, P.; Segaran, R.R.; Koh, L.P. Deep learning for environmental conservation. *Curr. Biol.* **2019**, *29*, R977–R982. Available online: https://www.sciencedirect.com/science/article/pii/S0960982219310322#bib21 (accessed on 10 March 2021). [CrossRef]
55. Kamminga, J.; Ayele, E.; Meratnia, N.; Havinga, P. Poaching detection technologies—A survey. *Sensors* **2018**, *18*, 1474. Available online: https://www.mdpi.com/1424-8220/18/5/1474/pdf (accessed on 10 March 2021). [CrossRef]
56. Yampolskiy, R.V. Personal universes: A solution to the multi-agent value alignment problem. *arXiv* **2019**, arXiv:1901.01851.
57. Allen, C.; Smit, I.; Wallach, W. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics Inf. Technol.* **2005**, *7*, 149–155. Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.307.7558&rep=rep1&type=pdf (accessed on 10 March 2021). [CrossRef]
58. United Nations, General Assembly. Universal Declaration of Human Rights A/RES/3/217A. 1948. Available online: https://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_RES_217(III).pdf (accessed on 11 April 2021).
59. Ruby, M.B. Vegetarianism. A blossoming field of study. *Appetite* **2012**, *58*, 141–150. Available online: https://vegstudies.univie.ac.at/fileadmin/user_upload/p_foodethik/Ruby__Matthew_2012._Research_Review_Vegetarianism._A_blossoming_field_of_study.pdf (accessed on 10 March 2021). [CrossRef]
60. IPBES. *Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*; Brondizio, E.S., Settele, J., Díaz, S., Ngo, H.T., Eds.; IPBES Secretariat: Bonn, Germany, 2019. Available online: https://ipbes.net/global-assessment (accessed on 10 March 2021).

61. Kahneman, D. *Thinking, Fast and Slow*; Farrar, Straus, & Giroux: New York, NY, USA, 2011.
62. McDevitt, M.A.; Dunn, R.M.; Spetch, M.L.; Ludvig, E.A. When good news leads to bad choices. *J. Exp. Anal. Behav.* **2016**, *105*, 23–40. Available online: https://core.ac.uk/download/pdf/208891271.pdf (accessed on 10 March 2021). [CrossRef]
63. Tomasik, B. Should We Intervene in Nature? *Essays on Reducing Suffering*. 2016. Available online: https://reducing-suffering.org/should-we-intervene-in-nature/ (accessed on 10 March 2021).
64. Horta, O. Debunking the Idyllic View of Natural Processes: Population Dynamics and Suffering in the Wild. *Télos* **2010**, *17*. Available online: https://www.academia.edu/2290959/Debunking_the_Idyllic_View_of_Natural_Processes_Population_Dynamics_and_Suffering_in_the_Wild (accessed on 10 March 2021).
65. Simmons, A. Animals, predators, the right to life, and the duty to save lives. *Ethics Environ.* **2009**, *14*, 15–27. [CrossRef]
66. Pearce, D. The Abolitionist Project. 2007. Available online: https://www.abolitionist.com/ (accessed on 10 March 2021).
67. Pearce, D. Compassionate biology. 2020. Available online: https://www.gene-drives.com/ (accessed on 10 March 2021).
68. Mak, K.K.; Pichika, M.R. Artificial intelligence in drug development: Present status and future prospects. *Drug Discov. Today* **2019**, *24*, 773–780. [CrossRef]
69. Burki, T. A new paradigm for drug development. *Lancet Digit. Health* **2020**, *2*, e226. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7194950/ (accessed on 10 March 2021). [CrossRef]