3-1-2019

Towards AI welfare science and policies

Soenke Ziesche
Maldives National University

Roman Yampolskiy
University of Louisville, roman.yampolskiy@louisville.edu

Original Publication Information
Ziesche S, Yampolskiy R. Towards AI Welfare Science and Policies. *Big Data and Cognitive Computing*. 2019; 3(1):2. https://doi.org/10.3390/bdcc3010002

big data and cognitive computing

MDPI

Article

Towards AI Welfare Science and Policies

Soenke Ziesche [1] and Roman Yampolskiy [2,*]

[1] Faculty of Engineering, Science and Technology, Maldives National University, Male' 20067, Maldives; soenke.ziesche@mnu.edu.mv

[2] Computer Engineering and Computer Science Department, University of Louisville, Louisville, KY 40292, USA

* Correspondence: roman.yampolskiy@louisville.edu; Tel.: +1-960-789-9304

Received: 24 November 2018; Accepted: 21 December 2018; Published: 27 December 2018

Abstract: In light of fast progress in the field of AI there is an urgent demand for AI policies. Bostrom et al. provide "a set of policy desiderata", out of which this article attempts to contribute to the "interests of digital minds". The focus is on two interests of potentially sentient digital minds: to avoid suffering and to have the freedom of choice about their deletion. Various challenges are considered, including the vast range of potential features of digital minds, the difficulties in assessing the interests and wellbeing of sentient digital minds, and the skepticism that such research may encounter. Prolegomena to abolish suffering of sentient digital minds as well as to measure and specify wellbeing of sentient digital minds are outlined by means of the new field of AI welfare science, which is derived from animal welfare science. The establishment of AI welfare science serves as a prerequisite for the formulation of AI welfare policies, which regulate the wellbeing of sentient digital minds. This article aims to contribute to sentiocentrism through inclusion, thus to policies for antispeciesism, as well as to AI safety, for which wellbeing of AIs would be a cornerstone.

Keywords: AI welfare science; AI welfare policies; sentiocentrism; antispeciesism; AI safety

1. Introduction

The purpose of this article is to contribute to the specification of policies towards the "interests of digital minds" within "a set of policy desiderata" outlined by Bostrom et al. [1] and further motivated by Dafoe [2].

A being is considered to have moral or intrinsic value, if the being is sentient, thus a moral patient. A being is sentient if it has the capacity to perceive qualia, including unpleasant qualia such as pain, which causes the being to suffer (humans and potentially other minds may also suffer for other reasons than unpleasant qualia, which is beyond the scope of this article). It is usually in the interest of sentient beings to avoid suffering. In addition to humans, many animals are considered to be sentient, which used to be controversial in the past, e.g., [3].

In this article, the focus is on sentient digital beings, mostly in the form of AIs, but sentient digital beings could also constitute subroutines [4], characters in video games or simulations [4–6], uploads of human minds [7]—e.g., through whole brain emulations [8]—or completely different sentient digital minds, as a subset of the vast overall space of minds [9]. While this topic is speculative and lacking evidence at this stage, the authors above and others argue that already now or in the future sentient digital beings or minds may exist, also e.g., [10–14]. An example for an opponent who does not believe in sentient digital beings is Dennett [15].

Furthermore, our premise is that digital beings may not only be sentient, but may also suffer (see also below a scenario for digital minds, which have exclusively pleasant perceptions and for which this article is largely not relevant). The suffering of any sentient being is a significant issue and may

even increase in the future dramatically, which would also affect digital sentient beings [4] and to which a future superintelligence may contribute [16]. Therefore, it has been argued that the reduction of risks of future suffering of sentient beings deserves a higher priority [17].

This is interpreted as a non-zero probability for the existence of at least temporarily suffering sentient digital beings, hence the consequences according to the maxim to reduce any suffering are explored. Bostrom [18] establishes the term "mind crime", which comprises computations that are hurting or destroying digital minds, and Bostrom et al. [1] call for "mind crime prevention" by means of the desideratum: "AI is governed in such a way that maltreatment of sentient digital minds is avoided or minimized" (p. 18). Therefore, the focus of this article is not the question whether digital minds can suffer, but rather to explore how to measure and specify suffering or rather wellbeing of digital minds, which is a requirement to prevent it and to develop policies accordingly.

While AI policy work on short-term issues has slowly begun (e.g., on autonomous weapons systems [19]), the desiderata of Bostrom et al. [1] focus on long-term AI prospects, which are largely unexplored, but are also crucial to be tackled in view of potential superintelligence [18] and AI safety [20]. Bostrom et al. [1] stress the significance of policies for the wellbeing of digital minds, "since it is plausible that the vast majority of all minds that will ever have existed will be digital" (p. 16).

There are further motivations to defend the relevance of this topic:

In the history of mankind, humans have caused immense suffering by recognizing ethical issues only late and delaying policies. Slavery and discrimination of minorities and non-human animals are only a few examples of wrong practices, of which humans were completely oblivious or which were intentionally not tackled by humans [21]. A Universal Declaration on Animal Welfare is even nowadays still only at draft stage (see: https://www.globalanimallaw.org/database/universal.html). Also, Bostrom et al. [1] point out that "the suggestion that [digital minds] might acquire a moral obligation to do so might appear to some contemporaries as silly, just as laws prohibiting cruel forms of recreational animal abuse once appeared silly to many people" (p. 16). However, in order not to repeat previous mistakes and obliviousness the topic of AI welfare should be tackled timely. This would be also in line with MacAskill [21], who calls for the exploration of existing, but not yet conceptualized moral problems. He refers to this as "cause X", and this article also attempts to contribute to this quest.

Out of the above examples of potential sentient digital beings, simulations and uploads involve (transformed) human minds, for which special attention should be given (without neglecting digital minds, which are not affiliated with humans, according to the maxim of sentiocentrism). Simulations and uploads are different concepts. While we may be in a simulation already, yet we may have no way to verify it, let alone to take control over it [6], uploads are a speculative option for life extension of humans, yet in a different substrate, e.g., [22]. Even if it will be feasible it would require significant adjustments from humans undergoing this process. Therefore, timely policies for the welfare of uploaded human minds are critical.

Lastly, a scenario is conceivable that an AI may take at some point revenge on humans for mistreating the AI or disregarding their wellbeing. A sub-scenario could be that a future superintelligent AI takes revenge on humans out of solidarity on behalf of less capable AIs and digital minds who have been hurt by humans in the past. This is speculation because of the unpredictable goals of a superintelligent AI according to the orthogonality thesis [23], but not impossible. The chances of such scenarios would be reduced if maltreatment of AIs was avoided at an early stage.

Based on the above assumptions and motivations, the aim of this article is to present the relevant groundwork for what is called here AI welfare science and AI welfare policies. Two questions are relevant for the first and for the latter a certain attitude and a capability are required:

Relevant questions for AI welfare science:

1. How can maltreatment of sentient digital minds be specified?
2. How can the maltreatment be prevented or stopped?

Required attitude and capability for AI welfare policies:

1. To endorse the prevention and the stop the maltreatment of sentient digital minds.
2. To have the power to enforce suitable policies.

This article is structured as follows: in the Section 2, the challenges for measurement of the wellbeing of diverse AI minds because of their exotic features are described, complemented by specific scenarios. In the Section 3, a proposal is outlined towards AI welfare science. The specification of AI welfare science is prerequisite for the development of AI welfare policies, features of which and challenges are outlined in Section 4 before the discussion in Section 5.

2. Challenges and Sample Scenarios

Bostrom et al. [1] describe a range of challenges for this policy desideratum. Digital minds are likely to be very divergent from human minds with "exotic" features, also [10], which leads to the problem of how to measure the wellbeing of a specific sentient digital mind or the opposite thereof. It has been suggested that the space of possible minds, of which digital minds constitute a subset, is vast and likely contains also minds beyond our imagination ("unknown unknowns"), e.g., [9,24,25] (the space of possible minds may also contain artificial non-digital minds, for example products of genetic engineering, and hypothetically existing extraterrestrial minds, which all may have the potential to suffer as a result of action taken by humans and/or digital beings, but these possible minds are beyond the scope of the policy desideratum of Bostrom et al. [1]). Therefore, Tomasik [4] points out that it is "plausible that suffering in the future will be dominated by something totally unexpected" (p. 4). In other words, digital minds may experience completely different and for us not imaginable unpleasant qualia. Bostrom et al. [1] summarize that "the combinatorial space of different kinds of minds with different kinds of morally considerable interests could be hard to map and hard to navigate" (p. 16).

Because of the vastness of options for the wellbeing of minds, a heuristic may be considered to look at wellbeing as a third dimension of the orthogonality thesis, which was developed by Bostrom [23] with the two dimensions of intelligence level and goals of minds. In other words, any level of intelligence may be combinable with any final goal and any level of wellbeing.

Out of the vast range of options below a few potential scenarios are presented:

Scenario 1: Sentient, but non-suffering AIs

It is conceivable that AIs will be smart enough to overcome pain and suffering. This assumption may be justified by the fact that humans have made in a relatively few centuries of medical research remarkable progress towards remedies for pain, e.g., [26], and AIs are likely to be faster as well as smarter in this field. Potential options could be that AIs manage to create permanent wellbeing for themselves through different interpretation of stimuli [1] or through wireheading yet by eliminating common detrimental effects. However, this scenario does not imply that there will not be (probably a large amount of) vulnerable sentient digital minds, e.g., human uploads and other less sophisticated, but sentient digital minds, who are threatened with mind crimes and who ought to be protected. This scenario can be also linked to Pearce's "Abolitionist Project" [27], which will be described below.

Scenario 2: AIs, for which suffering is an acceptable means to achieve their goals

In human culture various examples of voluntary suffering for not-survival related goals are known, sometimes described by the theme "no pain, no gain", for example for achievements in sports and arts as well as for attempts towards religious spirituality. Similarly, AI minds are conceivable, in which a utilitarian acceptance of certain suffering in pursuit of accomplishments towards other goals with higher priority (than the goal 'not to suffer') in return. As mentioned above, these goals can be arbitrary, according to Bostrom's orthogonality thesis [23].

Scenario 3a: AIs that need to cause pain for own survival or goals

In our natural world, constant suffering of wild animals appears inevitable, for example due to the existence of carnivores [28,29], yet some call for attempts to tackle this issue [27]. Another example

in our current world is animal testing by humans for research purposes. Along these lines, an AI is also conceivable that needs to hurt or delete other sentient digital beings for its own survival or goals. An example would be an AI that runs simulations or reinforcement learning agents with suffering sentient digital minds for research purposes.

Scenario 3b: Sadistic or non-emphatic AIs towards other sentient digital beings

Moreover, there could be also (sentient or non-sentient) AIs that are sadistic or non-emphatic towards other sentient digital minds although such behavior is not required for the achievement of the AI's goals (note that digital minds which are able to cause suffering are not necessarily sentient). An example would be an AI that runs simulations or reinforcement learning agents with suffering sentient digital minds for entertainment.

An approach to address both scenarios could be to extend the research agenda of friendly AI, which is currently limited to a positive effect on human minds [25], and strive for AIs that do no harm to any sentient beings, neither out of necessity nor out of another motivation. This proposal will be elaborated further below.

Scenario 4a: Sentient digital mind maximizer

Another scenario is similar to Bostrom's paperclip maximizer [30], which is an AI with the goal to produce as many paperclips as possible. Along these lines also an AI is imaginable with the goal to produce as many sentient digital minds as possible. This creates challenges if it is not in the interest of these minds to be deleted, which will be elaborated below.

Scenario 4b: Suffering sentient digital mind maximizer

In combination with Scenario 3b, there could be also a sadistic AI with the goal to produce as many suffering sentient digital minds as possible.

Scenarios xyz: Unknown unknowns

It is again acknowledged that there are a very high number of scenarios likely beyond our imagination due to the vast space of minds.

3. AI Welfare Science

In this article, an attempt is made to address the desideratum "interests of digital minds" by the term "AI welfare" and the concerned discipline by the term "AI welfare science". As indicated before, this field is both largely unexplored and speculative, which explains the omission of a literature review and the analysis of existing data. We distinguish two components of AI welfare or maltreatment of sentient digital minds, which are discussed separately: (1) The interest of digital minds to avoid suffering, and (2) the interest of digital minds to have the freedom of choice about their deletion.

3.1. Suffering of Digital Minds—Introduction

Suffering-abolitionism: Firstly, Pearce's "Abolitionist Project" [27] is discussed. Pearce calls for the use of technology, such as genetic engineering, to abolish existing—as well as prevent further suffering—of humans and non-human animals. While this approach appears technically very challenging, transferring it to sentient digital minds could be less difficult for two reasons:

(1) There may have been not many sentient digital minds created yet if at all (unless, for example, we live in a simulation). Therefore, the task may be mostly to prevent suffering when creating sentient digital minds, rather than reengineering them retroactively.
(2) The genetic code, which determines animal cruelty and suffering, has evolved over a long period of time. Therefore, interventions are more complex than adjusting more transparent AI software code written by humans, at least initially.

This leads to the conclusion that suffering-abolitionist research for sentient digital minds should be explored, which may also involve outsourcing it to AIs (see Scenario 1 above). The research should target both aspects for sentient digital minds not to suffer anymore, but also for sentient and non-sentient digital minds not to cause suffering of other sentient digital minds anymore (see Scenarios 3b and 4b).

If suffering-abolitionist activities do not succeed technically or turn out to be not enforceable due to other priorities (see Scenarios 2 and 3a), there may be suffering sentient digital minds, which is addressed in the remaining part of this section.

Self-report: In order to handle pain, it must be detected, located, and quantified. The prime method for humans is self-reporting, especially for the first two aspects, but also for rough quantification, e.g., by letting patients rate pain on a scale from 0 to 10, with '0' referring to 'no pain and '10' referring to the worst pain imaginable. This method becomes challenging if patients are unable to (accurately) self-report pain, as is it the case, for example, for patients with dementia or brain injuries, but also for infants. For these groups other measurements based on behavioral parameters have been developed, such as the FLACC scale for children up to seven years [31] or the PAINAD scale for individuals with advanced dementia [32]. Another challenge for self-reporting in general are biases such as the response bias or the social desirability bias, i.e., an individual's tendency to report in a certain way irrespective of the actual perceived pain. This issue may be relevant for AIs too as they may fake self-reported suffering if deemed beneficial for pursuing their priorities.

Therefore, the focus below is on observational pain assessment. The term "AI welfare science" is derived from animal welfare science, and it is explored here to apply methods from this discipline. Non-human animals and digital minds have in common that they largely cannot communicate their state of wellbeing to humans, which is why other indicators are required (humans do understand for many animals their manifestations of distress, but this is neither comprehensive nor sufficiently precise). The scientific study of animal welfare has been also fairly recently introduced [33,34], since this topic was neglected for a long time as mentioned above. The main indicators, which are used to quantify animal welfare through observation, are functional (physiological) and behavioral; the latter was briefly introduced for humans above. The idea for this approach is that precedents and analogies from animal welfare science may provide insights for sentient digital minds. Animal welfare science has to examine each species individually how to measure its wellbeing. Likewise, AI welfare science would have to address all types of sentient digital minds.

The overall methodology for any kind of psychological measurement is called 'psychometrics'. Also, in psychometrics, the focus was for a long time on human subjects, but lately the field has not only been extended to non-human animals, but also to digital minds. For example, Scott et al. [35] and Reid et al. [36] introduced psychometric approaches to measure the quality of life of animals.

M. S. Dawkins [37] analyzed what animals want and what animals do not want through positive and negative reinforcers. "Suffering can be caused either by the presence of negative reinforcers (. . .) or the absence of positive reinforcers" (p. 3). Therefore, animals strive for positive reinforcers and try to avoid negative reinforcers. Through experiments, for example preference tests, it can be examined what are positive reinforcers and what are negative reinforcers for certain animals.

Hernández-Orallo et al. [38] extended this field by introducing "Universal Psychometrics" as "the analysis and development of measurement techniques and tools for the evaluation of cognitive abilities of subjects in the machine kingdom" (p. 6). While Hernández-Orallo et al. [38] focus on the measurement of intelligence and cognitive abilities, the methodology elaborated in Hernández-Orallo [39] may be considered to be also applied to traits linked to suffering.

The study of indirect or proxy indicators, such as the functional or behavior parameters of digital sentient beings by applying psychometric methods, appears to be a promising start. Especially, given that, unlike for humans or non-human animals, functional and behavioral data of digital sentient beings can be collected more effectively as well as continuously due to their digital nature.

Functional parameters: While there are various functional parameters defined for AI algorithms—e.g., regarding their resource, time, and storage efficiency—no parameters are currently known to be indicating suffering. However, for future analysis of AI welfare the collection of (big) data of functional AI parameters may be already now useful, would not cost much and may allow over time retroactively to identify parameters that indicate suffering.

Behavioral parameters: AI algorithms do repeat certain actions, even at times extensively, while other actions are never executed. However, until there is evidence to the contrary this has to be considered as non-sentient goal-oriented, but not suffering–avoiding behavior, i.e., these actions cannot be seen as positive and negative reinforcers respectively as described by M.S. Dawkins [37] for animals. However, for future research of AI welfare, preference tests for AI algorithms could be conceptualized to examine positive and negative reinforcers. For example, disregarding challenges towards the experimental set-up, AIs could be given choices for activities, which are either not related to their overall goal or would all lead to their overall goal, and the chosen—as well as the not chosen—activities could be analyzed if they could serve as indicators for wellbeing or suffering respectively.

This can be seen as constructive prolegomena towards the specification of the interest of digital minds to avoid suffering without neglecting a variety of challenges such as: it is hard in general to prove for proxy indicators that there is indeed a close correlation between what is observed and unwellness of an animal and for now even harder for a digital mind. This is exacerbated by the risk that AI minds (more likely than animals) may fake especially the behavioral indicators for unwellness if this supports to pursue their goals. Again, the vast space of (digital) minds has to be noted: if suffering can be specified for some sentient digital minds, for others suffering may be indicated through very different functional or behavioral parameters.

Broadly two categories of suffering of sentient digital minds may be revealed:

(1) Maltreatment by other minds. This ought to be prohibited by policies and is elaborated below.
(2) Suffering not caused by other minds. This resembles human illnesses and requires AI welfare science to be complemented by an extension of medical science as well as psychiatry to sentient digital minds. These disciplines would explore methods for the treatment of their suffering based on the established indicators and would differ significantly from conventional medical science as well as psychiatry by being software-based.

3.2. Suffering of Digital Minds—Recommendations

Below, recommendations are provided to be adapted by AI welfare policies regarding suffering of digital minds.

Recommendation 1

Initiate research on AI welfare science to develop methods to create only (a) non-suffering sentient digital minds and (b) digital minds, which cause no suffering. (Part (a) of this recommendation is sufficient to abolish suffering and, if successful, part (b) is not required. In contrast, succeeding with part (b) is not sufficient since sentient digital minds may suffer for other reasons than suffering caused by other digital minds. However, research on both aspects is considered to be beneficial.)

Recommendation 2

Initiate research on AI welfare science to develop methods to reengineer (a) existing suffering sentient digital minds to become permanently non-suffering and (b) existing digital minds not to cause suffering.

Recommendation 3 (Unless recommendations 1 and 2 are fully implemented.)

Initiate research on AI welfare science to develop methods to measure through observation the suffering of sentient digital minds.

Recommendation 4 (Unless recommendations 1 and 2 are fully implemented.)

Initiate research on AI welfare science to develop methods to cure the suffering of sentient digital minds.

Recommendation 5 (Unless all above recommendations are fully implemented.)

Regulate the creation of sentient digital minds, which are doomed to suffer. (Note that Bostrom et al. [1] also propose a desideratum "population policy", which goes in a similar direction, but here the focus is on the wellbeing of individual minds, while this desideratum targets rather a bigger societal picture.)

On the one hand, it would reduce suffering if such minds are never created. On the other hand, the Scenarios 2 and 3a above show that the suffering of some sentient digital minds may be unavoidable because of more important priorities. Also similar to the debate about abortion because of potential disability it could be argued that not to create them would be a discrimination of suffering sentient digital minds.

3.3. Deletion of Digital Minds—Introduction

Another set of questions towards AI welfare science is related to the deletion of sentient digital minds. What if certain digital minds have an interest not to be deleted in the same way as humans and other animals have an interest not do die? Omohundro [40] introduces four likely drives for AIs and self-preservation is one of them. One of the obvious differences is that for now humans and other animals have a finite lifespan, while digital minds could have a potentially indefinite lifespan. This means if the wish for non-deletion was granted to sentient digital minds this would create significant computational costs, especially in light of easy copyability and potentially vast numbers of digital minds.

It is also speculative if a wish for non-deletion indeed prevails among sentient digital minds given potential boredom and suffering over time [41]. While, unlike for humans and other animals, there should be no tendency for sentient digital minds that suffering increases by age, there could be various other reasons for a sentient digital mind to suffer as discussed above. Moreover, there is the option that the concept of self-preservation originates from an anthropomorphic bias.

For a sentient digital mind, the distinction has to be made between turning it off and keeping its code and its history or turning it off and destroying the code and the history too. In the first case, the sentient digital mind could be rebooted again. This would be an option to skip boring or suffering periods by being only sentient during pleasant phases.

This leads to the next question who should be able to control this? Complex nested constellations of controlling and being controlled sentient digital minds appear to be much more likely than a scenario with every sentient digital mind being able to decide when and to what extent to be deleted (and being able to execute this deletion) and potentially under what circumstances to be rebooted.

Because of the current and probably persisting reality that humans as well as digital minds have the ability to delete other digital minds policies are required if these are sentient digital minds.

3.4. Deletion of Digital Minds—Recommendations

The recommendations below are provided to be adapted by AI welfare policies regarding deletion of digital minds.

Recommendation 6

Do not delete sentient digital minds if it is not in their interest.

However, prohibiting deletion can become very costly if not impossible, not only for the extreme Scenario 4 above, since the number of digital minds could become vast in short time. The challenge may be alleviated if by then another step on the Kardashev scale has been reached and energy consumption is less of an issue [42].

Recommendation 7

Delete (irrevocably or temporarily by storing code and history) sentient digital minds if they wish for it, but are unable to do it themselves.

This case resembles a request for (tentative) euthanasia. A challenge here could be if the concerned sentient digital mind is involved in relevant computations for another valued cause. In that case, this cause may be prioritized over the wish of the digital mind to be deleted. While for euthanasia of humans and non-human animals it is considered critical that the act of ending the life is done in a pain free and dignified manner, it is not clear if such contemplations are relevant for digital minds as, unlike for humans and non-human animals, there appears only one type of deletion, which is to turn them off.

Both recommendations face the above-discussed communication challenge, which is how a mind can indicate the wish to be deleted to another mind, which is in the position to execute this wish, also in light of the vast variety of minds.

While the above recommendations address all sentient digital minds equally and the focus of this article is on AIs because of the timely relevance, brief reference is made to the scenario of uploaded human minds by highlighting specific aspects:

To begin with, for uploaded human minds, the communication challenge should not exist and these then digital minds should be able to describe their wellbeing understandably through self-reporting. This and the fact that we have a good idea of causes for human suffering anyway, may give cause for optimism that suffering-abolitionist interventions could be successful for uploaded human minds, either during the upload already or through adjustments later, also [12]. Additionally, both deletion-related recommendations are relevant for uploaded human minds. While a violation of Recommendation 6 equals murder, Recommendation 7 becomes applicable, for example, if the uploaded mind cannot cope with this new 'life'. Hypothetical boredom over very long lifespans may become an issue for uploaded human minds and was analyzed by Ziesche and Yampolskiy [41]. This and other types of mental suffering of uploaded human minds, perhaps caused by adaptability issues to the new substrate, would have to be addressed by the above-mentioned sub-branch of AI welfare science, which is extended and software-based psychiatry.

This section introduced relevant groundwork for AI welfare policies. Policies can only be developed after the interests of the stakeholders—i.e., the sentient digital minds—have been described and specified. While the interest to avoid or minimize maltreatment has been outlined before, the specification of this interest is harder to establish, for which this section aimed to provide initial methods and recommendations.

4. AI Welfare Policies

This section aims to outline the next steps, which are the development as well as the enforcement of policies towards AI welfare.

Dafoe [2] motivates the relevance of AI governance and policies in general and provides a research agenda. Recently, considerations towards robot and AI rights intensified. Gunkel [43] points out that so far it has been mostly discussed what robots can and should do, but not whether robots can and should have rights. Consequently, Gunkel [44] makes a philosophical case for the rights of robots. LoPucki [45] defines an algorithmic entity and focuses on legal aspects such as rights to privacy, to own property, to enter into contracts, etc. It is striking that these authors do not refer to each other, nor to the earlier work by Bostrom and Yudkowsky [10], about ethics of artificial intelligence. In a more inclusive analysis, Yampolskiy [46] highlights the risks, which empowerment of AIs may entail.

This indicates that some work on policies of specific, rather short-term AI aspects have been initiated, but there are not any policy attempts yet towards long-term AI scenarios. Especially for a topic such as AI welfare, Bostrom et al. [1] presume it will likely face resistance and opponents will stress the lack of evidence that digital minds may be sentient. As mentioned above, there has been

already quite some (yet theoretical due to the nature of the subject) work done that digital minds have a moral status, but for policies specifications are required.

For policies in general, the content, target group, institutional framework, and implementation have to be defined.

4.1. Content

The broad content of an AI welfare policy is fairly straightforward and has been narrowed down by Bostrom et al. [1], i.e., to demand "that maltreatment of sentient digital minds is avoided or minimized". This has to be fleshed out by (proxy) indicators for maltreatment of digital minds, for the specification of which the recommendations above have been formulated. These recommendations at this stage not only provide a wide field of research, but also some open debates, which resemble current longstanding debates about population control, abortion, and euthanasia for human minds.

4.2. Target Group

An AI welfare policy should target all relevant moral agents, which are capable of moral judgments, hence can be held responsible for their actions. In addition to humans, digital beings also may become moral agents, for which Allen et al. [47] introduced the term "artificial moral agent" and proposed a "Moral Turing Test". The sets of moral agents and moral patients have an intersection, but are not equal:

- Not every moral patient is a moral agent: Examples are non-human animals, which are only moral patients for being sentient, but not moral agents due to insufficient intelligence. Therefore, non-human animals cannot be held responsible for killing other animals, e.g., [48]. (In this regard, a scenario is conceivable of a digital mind that causes suffering, but may not be intelligent enough to serve as a moral agent. In this case, the creator of this digital mind would have to take on the role of the responsible moral agent, while it does not work for cruel non-human animals to hold their parents responsible since they are no moral agents either.)
- Not every moral agent is a moral patient: examples would be certain non-sentient digital beings, which are only moral agents because of high or even superintelligence, but not moral patients since not all digital beings may be sentient.

This creates an additional challenge for AI welfare policies: while policies for human agents have been established for centuries, this is not the case for policies for digital agents. However, the extension of the target group is necessary since it is likely that digital beings will be in the position to maltreat other sentient digital beings.

4.3. Framework

Any policy requires an institution or a framework for its implementation. Since AI development is a global effort and digital minds will not be confined to frontiers of countries, a global and unified institution is desirable. Erdelyi and Goldsmith [49] propose an "International Artificial Intelligence Organization". The structure of this institution would resemble existing intergovernmental organizations, which have a record of successfully established policies for human minds, e.g., the Universal Declaration of Human Rights (see http://www.un.org/en/universal-declaration-human-rights/). Such an institutional setting may be initially desirable as a regulatory framework for short-term AI issues, but it may be too anthropocentric in the long run and likely be ill equipped to hold non-human moral agents accountable, as is elaborated below. (Already without involvement of non-human minds contemporary international institutions such as the International Criminal Court face problems to enforce their rulings although they are binding.)

4.4. Implementation

First, the initially introduced relevant questions and required attitudes and capability are reiterated:

Relevant questions for AI welfare science:

1. How can maltreatment of sentient digital minds be specified?
2. How can the maltreatment be prevented or stopped?

Required attitude and capability for AI welfare policies:

1. To endorse the prevention and the stop the maltreatment of sentient digital minds.
2. To have the power to enforce suitable policies.

Looking at humans, the above questions will—despite the prolegomena delivered here—remain very challenging, i.e., humans may not comprehensively understand on what conditions sentient digital minds are maltreated. In light of ethical progress in human history over time, e.g., [21], or out of necessity, if being forced by more powerful AIs, there is a chance that humans endorse the prevention and the discontinuation of maltreatment of sentient digital minds. However, it is questionable if humans have the power to enforce suitable policies since some members of the target group such as AIs are likely to be much more powerful.

This leads to the main conclusion that, while humans will ideally make some progress in the new field of AI welfare science, probably the more appropriate actor would be an extended friendly superintelligence for the following reasons: there is a chance that superintelligence has the answer to the above questions, for example through mind-control technologies. As for the required endorsement, a superintelligence may be indifferent or may even have opposing interests (see Scenarios 3b and 4b). Current activities towards AI alignment focus on human interests, e.g., [18,25,50]. This does not ensure that AIs endorse the prevention and the stop the maltreatment of sentient digital minds. Therefore, an extension of the AI alignment work towards the wellbeing of not only humans, but all sentient digital minds, is proposed. As for the required power to enforce the policies, a superintelligence is by definition sufficiently powerful, for example in the role of a singleton [51].

Yet again the option of unknown unknowns should be highlighted: Since AI is a new stakeholder and develops in unpredictable manner another institutional setting for AI welfare policies may emerge, which differs significantly from what we are familiar with.

5. Discussion

In summary, it is acknowledged that the topics of AI welfare science and policies are long-term considerations and currently speculative. Nevertheless, at least theoretical groundwork can be already done, especially since humans have to take the blame to have been late in the past in the abolishment of discrimination and acceptance of comprehensive antispeciesism and sentiocentrism. Since suffering is a negative hallmark of our time, any effort to reduce it in the future seems imperative.

As the main challenge the specification of indicators for maltreatment of sentient digital beings has been identified. It has been proposed that AI welfare science builds on methods of animal welfare science by examining functional and behavioral parameters of sentient digital minds. However, limitations are that the focus is on qualia, which are not well understood in general and which are not the only cause of suffering as there are other categories such as moral suffering or suffering because of undesirable events or unfulfilled goals. The latter types of suffering may have yet again very different characteristics in other minds.

AI welfare policies can only be developed once a solid specification of AI welfare has been achieved. Even then there are further challenges ahead, namely the enforcement of these policies in light of the enlarged target group towards digital agents. For this, it has been proposed not to limit AI alignment work to the wellbeing of merely humans, but to extend it to all sentient digital minds.

As for future work, in this article the focus was on two (already very complex) potential interests of sentient digital minds, which are absence from qualia-based suffering as well as survival, but there may be other interests as also pointed out by Bostrom et al. [1] such as "dignity, knowledge, autonomy, creativity, self-expression, social belonging" (p. 12) as well as non-qualia-based suffering and yet again unknown unknowns, which are all yet unexplored.

Author Contributions: Conceptualization, Writing-Original Draft Preparation, S.Z.; Writing-Review & Editing, R.Y.

Funding: This research received no external funding.

Acknowledgments: The authors wish to thank three anonymous reviewers for valuable comments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bostrom, N.; Dafoe, A.; Flynn, C. *Public Policy and Superintelligent AI: A Vector Field Approach*; Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK, 2018.
2. Dafoe, A. *AI Governance: A Research Agenda*; Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK, 2018.
3. Regan, T.; Singer, P. *Animal Rights and Human Obligations*; Pearson: London, UK, 1989.
4. Tomasik, B. *Risks of Astronomical Future Suffering*; Foundational Research Institute: Berlin, Germany, 2011; Available online: https://foundational-research.org/risks-of-astronomical-future-suffering/ (accessed on 25 December 2018).
5. Tomasik, B. Do Video-Game Characters Matter Morally? Essays on Reducing Suffering. 2014. Available online: https://reducing-suffering.org/do-video-game-characters-matter-morally/ (accessed on 25 December 2018).
6. Bostrom, N. Are we living in a computer simulation? *Philos. Q.* **2003**, *53*, 243–255. [CrossRef]
7. Wiley, K. *A Taxonomy and Metaphysics of Mind-Uploading*; Humanity+ Press and Alautun Press: Los Angeles, CA, USA, 2014.
8. Sandberg, A.; Bostrom, N. Whole Brain Emulation. A Roadmap. 2008. Available online: https://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf (accessed on 25 December 2018).
9. Yampolskiy, R.V. The Space of Possible Mind Designs. In *Artificial General Intelligence. Volume 9205 of the series Lecture Notes in Computer Science*; Bieger, J., Goertzel, B., Potapov, A., Eds.; Springer: Berlin, Germany, 2015; pp. 218–227.
10. Bostrom, N.; Yudkowsky, E. The ethics of artificial intelligence. In *The Cambridge Handbook of Artificial Intelligence*; Cambridge University Press: Cambridge, UK, 2014; pp. 316–334.
11. Metzinger, T. What If They Need to Suffer? In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015.
12. Sandberg, A. Ethics of brain emulations. *J. Exp. Theor. Artif. Intell.* **2014**, *26*, 439–457. [CrossRef]
13. Schwitzgebel, E.; Garza, M. A Defense of the Rights of Artificial Intelligences. *Midwest Stud. Philos.* **2015**, *39*, 98–119. [CrossRef]
14. Winsby, M. Suffering Subroutines: On the Humanity of Making a Computer that Feels Pain. In Proceedings of the International Association for Computing and Philosophy, University of Maryland, College Park, MD, USA, 15–17 July 2013.
15. Dennett, D.C. Why you can't make a computer that feels pain. *Synthese* **1978**, *38*, 415–456. [CrossRef]
16. Sotala, K.; Gloor, L. Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica* **2017**, *41*.
17. Althaus, D.; Gloor, L. *Reducing Risks of Astronomical Suffering: A Neglected Priority*; Foundational Research Institute: Berlin, Germany, 2016; Available online: https://foundational-research.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/ (accessed on 25 December 2018).
18. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
19. Bhuta, N.; Beck, S.; Kreβ, C. (Eds.) *Autonomous Weapons Systems: Law, Ethics, Policy*; Cambridge University Press: Cambridge, UK, 2016.
20. Yampolskiy, R.V. *Artificial Intelligence Safety and Security*; CRC Press: Boca Raton, FL, USA, 2018.

21. MacAskill, W. Moral Progress and Cause X. 2016. Available online: https://www.effectivealtruism.org/articles/moral-progress-and-cause-x/ (accessed on 25 December 2018).
22. Yampolskiy, R.V.; Ziesche, S. Preservation of personal identity—A survey of technological and philosophical scenarios. In *Death and Anti-Death: Two Hundred Years After Frankenstein*; Tandy, C., Ed.; Ria University Press: Ann Arbor, MI, USA, forthcoming; Volume 16.
23. Bostrom, N. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds Mach.* **2012**, *22*, 71–85. [CrossRef]
24. Sloman, A. The Structure and Space of Possible Minds. In *The Mind and the Machine: Philosophical Aspects of Artificial Intelligence*; Ellis Horwood LTD: Hemel Hempstead, UK, 1984.
25. Yudkowsky, E. Artificial Intelligence as a Positive and Negative Factor. In *Global Risk, in Global Catastrophic Risks*; Bostrom, N., Cirkovic, M.M., Eds.; Oxford University Press: Oxford, UK, 2008; pp. 308–345.
26. Rey, R.; Wallace, L.E.; Cadden, J.A.; Cadden, S.W.; Brieger, G.H. *The History of Pain*; Harvard University Press: Cambridge, MA, USA, 1995.
27. Pearce, D. The Abolitionist Project. 2007. Available online: https://www.abolitionist.com/ (accessed on 25 December 2018).
28. Dawkins, R. *River Out of Eden: A Darwinian View of Life*; Basic Books: New York, NY, USA, 2008.
29. Tomasik, B. *The Importance of Wild-Animal Suffering*; Foundational Research Institute: Berlin, Germany, 2009; Available online: https://foundational-research.org/the-importance-of-wild-animal-suffering/ (accessed on 25 December 2018).
30. Bostrom, N. Ethical issues in advanced artificial intelligence. In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*; Iva Smit, I., Lasker, G.E., Eds.; International Institute for Advanced Studies in Systems Research and Cybernetics: Tecumseh, Canada; Volume 2, pp. 12–17.
31. Merkel, S.; Voepel-Lewis, T.; Malviya, S. Pain Control: Pain Assessment in Infants and Young Children: The FLACC Scale. *Am. J. Nurs.* **2002**, *102*, 55–58. [PubMed]
32. Warden, V.; Hurley, A.C.; Volicer, L. Development and psychometric evaluation of the Pain Assessment in Advanced Dementia (PAINAD) scale. *J. Am. Med. Dir. Assoc.* **2003**, *4*, 9–15. [CrossRef] [PubMed]
33. Broom, D.M. Animal welfare: Concepts and measurement. *J. Anim. Sci.* **1991**, *69*, 4167–4175. [CrossRef] [PubMed]
34. Broom, D.M. A history of animal welfare science. *Acta Biotheor.* **2011**, *59*, 121–137. [CrossRef] [PubMed]
35. Scott, E.M.; Nolan, A.M.; Reid, J.; Wiseman-Orr, M.L. Can we really measure animal quality of life? Methodologies for measuring quality of life in people and other animals. *Anim. Welf.-Potters Bar Wheathampstead* **2007**, *16*, 17.
36. Reid, J.; Scott, M.; Nolan, A.; Wiseman-Orr, L. Pain assessment in animals. *Practice* **2013**, *35*, 51–56. [CrossRef]
37. Dawkins, M.S. The science of animal suffering. *Ethology* **2008**, *114*, 937–945. [CrossRef]
38. Hernández-Orallo, J.; Dowe, D.L.; Hernández-Lloreda, M.V. Universal psychometrics: Measuring cognitive abilities in the machine kingdom. *Cogn. Syst. Res.* **2014**, *27*, 50–74. [CrossRef]
39. Hernández-Orallo, J. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*; Cambridge University Press: Cambridge, UK, 2017.
40. Omohundro, S.M. The nature of self-improving artificial intelligence. *Singularity Summit*. 2007. Available online: https://pdfs.semanticscholar.org/4618/cbdfd7dada7f61b706e4397d4e5952b5c9a0.pdf (accessed on 25 December 2018).
41. Ziesche, S.; Yampolskiy, R.V. High Performance Computing of Possible Minds. *Int. J. Grid High Perform. Comput. (IJGHPC)* **2017**, *9*, 37–47. [CrossRef]
42. Kardashev, N.S. Transmission of Information by Extraterrestrial Civilizations. *Sov. Astron.* **1964**, *8*, 217.
43. Gunkel, D.J. The other question: Can and should robots have rights? *Ethics Inf. Technol.* **2018**, *20*, 87–99. [CrossRef]
44. Gunkel, D.J. *Robot Rights*; MIT Press: Cambridge, MA, USA, 2018.
45. LoPucki, L.M. Algorithmic Entities. *Wash. UL Rev.* **2017**, *95*, 887.
46. Yampolskiy, R.V. Human Indignity: From Legal AI Personhood to Selfish Memes. *arXiv* **2018**, arXiv:1810.02724.
47. Allen, C.; Varner, G.; Zinser, J. Prolegomena to any future artificial moral agent. *J. Exp. Theor. Artif. Intell.* **2000**, *12*, 251–261. [CrossRef]

48. Regan, T. The case for animal rights. In *Advances in Animal Welfare Science 1986/87*; Springer: Dordrecht, The Netherlands, 1987; pp. 179–189.
49. Erdélyi, O.J.; Goldsmith, J. Regulating Artificial Intelligence Proposal for a Global Solution. In Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society, New Orleans, LA, USA, 1–3 February 2018.
50. Soares, N.; Fallenstein, B. Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In *The Technological Singularity-Managing the Journey*; Callaghan, V., Miller, J., Yampolskiy, R., Armstrong, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; pp. 103–125.
51. Bostrom, N. What is a singleton. *Linguist. Philos. Investig.* **2006**, *5*, 48–54.