AAAI: an Argument Against Artificial Intelligence

Sander Beckers
Department of Computer Science, Cornell University
Sander.Beckers@cornell.edu

Abstract

The ethical concerns regarding the successful development of an Artificial Intelligence have received a lot of attention lately, and rightly so. Even if we have good reason to believe that it is very unlikely, the mere possibility of an AI causing extreme human suffering is problematic enough to warrant serious consideration. In this paper I argue that a similar ethical concern arises when we look at this problem from the opposite perspective, namely that of the AI itself. Even if we have good reason to believe that it is very unlikely, the mere possibility of humanity causing extreme suffering to an AI is problematic enough to warrant serious consideration. I shall draw the conclusion that humanity should not attempt to create an AI.

Introduction

I here present an argument on moral grounds against the attempt to create an Artificial Intelligence, which occurred to me after having attended a very fascinating conference on Ethics and AI at NYU.[1] Besides the obvious issue of the potential negative impact that a future AI might have on humanity, many speakers also addressed concerns regarding the ethical impact on the AI's themselves. These concerns arise from the underlying assumption that if an AI reaches high levels of intelligence, both in terms of reasoning as in its capacity to consciously experience emotions, then it ought to be considered a moral agent.

My argument is a dramatic amplification of such concerns, based on the idea that not only could an AI reach human-like levels of intelligence and ethical awareness, but it might even acquire "superhuman" levels of these properties. The leap from the human-like to the superhuman was suggested by several of the speakers, which led me to believe that it has some plausibility.[2]

[1] https://wp.nyu.edu/consciousness/ethics-of-artificial-intelligence/

[2] The talk of Steve Petersen is the first that comes to mind, but also the talks given by Nick Bostrom, as well as those of Wendell Wallach, S. Matthew Liao, Eric Schwitzgebel and Mara Garza, and that of John Basl and Ronald Sandler, implied in some way or other that we cannot rule out the possibility that an AI would have capacities such that it deserves a special moral status. The program of the conference is available at: https://18798-presscdn-pagely.netdna-ssl.com/consciousness/wp-content/uploads/sites/2713/2016/10/FinalProgramEthicsofAI.pdf

I would like to stress that I am not at all convinced that an AI will develop such superhuman capacities, in fact I am not even convinced that the very notion of such capacities is meaningful. However I *am* convinced that the mere possibility of such capacities is entirely in line with much of what has been said at that conference, and in fact is implicit in many widespread views within the AI-community. Given the dramatic consequences that such capacities could produce, it is worthwhile exploring them.

In a nutshell, the argument is very simple: given that many AI researchers assume the possibility that an AI could experience an extreme form of suffering, it would be better if humanity avoids the risk of this occurring and does not attempt to build an AI. Further, I argue that the possible benefits of having an AI do not offer sufficient weight to undermine this conclusion. However, given the lack of control that any individual has over the actions of humanity in general, I conclude that anyone who takes this argument seriously *should nevertheless* attempt to build an AI.

The Argument for Supersuffering

First some terminology.

Definition 1. *A scenario is* possible *if the probability of it occurring is strictly positive.*

I take my first statement to be uncontroversial, as do the overwhelming majority of AI-researchers.

Premise 1. *If humanity attempts to create an AI, then it is possible that we will be successful.*

The notion of superintelligence is used quite regularly in the AI-community. More generally, several speakers of the above mentioned conference implicitly endorsed the view that many other human properties are also quantifiable, and could be developed by an AI to a degree that reaches far beyond anything that we have ever seen.

Definition 2. *For any property* $prop$ *which is both quantifiable and applies to human beings, a being has the property* superprop *if the being has* $prop$ *to a degree that is greater*

than the sum of the degrees for $prop$ of all human beings alive today.

The focus on the ethics of AI is usually discussed from the perspective of its behaviour towards human beings, rather than the other way around. The idea is that we should make sure that an AI is highly sensitive to our ethical values, to prevent it from radicalising its objectives in such a manner that it treats human beings as mere instruments. But if we succeed in making an AI that is capable of an extreme degree of empathy towards human beings, so that it is acutely sensitive to even the slightest suffering, then it seems entirely plausible that it is capable of an extreme form of suffering itself. I will focus on the superproperty of *supersuffering*, since it offers the clearest example of something which we have a moral duty to try and avoid.

Premise 2. *It is possible that a future AI has the capacity for supersuffering.*

Admittedly, my definition sets the bar quite high for achieving the level required for a superproperty. Nonetheless, once one has accepted the quantification of such notions to begin with, then it is hard to see how one can deny the assumption that it is at least *possible* for an AI to reach such high levels. After all, any of our current predictions on what an AI would look like involve massive uncertainty, and hence it would be unwise to straight out dismiss this possibility.

Further, for reasons of simplicity the present argument focuses on the suffering of a single AI. However it is reasonable to assume that if we are successful at creating an AI, we will produce a large number of copies. From there it's only a small step to imagine that any undesirable and unexpected capacity of an AI would be present in each of them. Given that such a capacity is unintentional, it might very well be that there exists a trigger for instantiating it in all of them at the same time. (One could imagine some complicated version of the millennium bug.) Therefore one can lower the threshold for supersuffering by several orders of magnitude if one switches focus to the *total* suffering of all AI's, rather than that of a single one.

Moreover there is the following sensible principle that was formulated by Bostrom and Yudowsky (2014, p. 326).

Premise 3 (Principle of Subjective Rate of Time)**.** *In cases where the duration of an experience is of basic normative significance, it is the experience's subjective duration that counts.*

Given the speed at which we can expect an AI to be operating, this principle in and of itself is already sufficient to highlight how the experience of suffering for an AI can take on far more extreme forms than it can for human beings: a single experiment that goes astray for a few seconds could result in an AI suffering for many years.

For example, assume we are running an experiment in order to fine-tune a specific parameter that controls how much pain the AI feels. So we simulate a million scenarios in which it might be hurt, and assess if it reacts accordingly. This implies that an AI could go through a million of experiences of pain during just a fraction of a second. Now even if we intended our experiment to minimise suffering by using a very low value for the pain parameter, a simple bug – like forgetting that a variable is a pointer if programming in C – could result in that parameter being accidentally set to a thousand times its intended value. In that case, the AI could suffer more pain in a single second than the total pain that has been suffered by all of humanity. (For a strikingly realistic demonstration of a similar scenario, see the episode "White Christmas" of the superb speculative fiction TV-series *Black Mirror*.)

For another illustration, one could imagine that the experience of empathy is achieved in the AI by replicating any suffering that it observes. Combined with the fact that the AI might have access to all of human history, it could suffer the entire extent of human suffering simply by going through its memory.

I take the next premise to be evident.

Premise 4. *If a being has the capacity for X, then it is possible that X will be instantiated.*

This brings me to my first conclusion, which if accepted would call for a serious reflection on the project of creating an AI.

Conclusion 1. *It is possible that an AI will suffer to an extent which is greater than all possible suffering of human beings alive today.*

If one accepts this conclusion, then in the very least a convincing argument is required to demonstrate that the possibility of supersuffering is an acceptable price to pay. I see three straightforward suggestions for doing so:

1. The expected benefits for mankind that come from creating an AI outweigh the possibility of supersuffering.
2. The attempt at creating an AI is not at all special in this regard, since all other acts that we perform as humanity today are also possible causes of extreme suffering in the future, and nevertheless we find this perfectly acceptable.
3. The negative scenario of supersuffering is compensated by a positive scenario of an AI experiencing superpleasure.

In the remainder of this paper the aim is to show how all three suggestions fail.

The Ethical Priority of Artificial Suffering

The possible benefits as well as the possible harms that an AI could bring to humanity obviously take on many forms. Hence I will assume a very tolerant notion of both pleasure and suffering, and simply ignore anything which is unsuited for quantification. In this manner we can quantify the expected benefits of an AI by way of the expected overall increase in human pleasure (which would be negative in case we expect there to be more suffering than pleasure).

Definition 3. *We denote by $E(HP)$ the expected overall increase in human pleasure that is the result of having created an AI. Further, $P(SS)$ denotes the probability that an AI will supersuffer, given that we have created one, and Suf denotes the degree of suffering that is minimally required for supersuffering.*

The next premise is almost a direct consequence of the Conclusion 1. All it does is make explicit that however small the probability of an AI supersuffering, it is not small enough to factor out the immense difference between the supersuffering of an AI compared to the pleasure of humanity. Therefore rejecting this premise (whilst accepting the above conclusion) would require such a precise estimate of said probability, that it is hard to see what type of evidence could support it.

Furthermore, it is only fair to assume that the abundance of arguments predicting the end of humanity – or its enslavement – if an AI is created reduces the expected pleasure of humanity $E(HP)$ significantly. Even the most ardent sceptic of such arguments should admit that the mere possibility of humanity going extinct or becoming enslaved by an AI is a pretty grim prospect, and thus it should have some weight in reducing the expected outcome.

Premise 5. $P(SS) * Suf >> E(HP)$.

Thus far I have not made explicit that an AI is a moral agent. I take for granted that any conscious and intelligent being which has human-like or superhuman capacities deserves moral consideration in the same way that a human does.

Premise 6. *When evaluating the overall expected benefits of creating an AI, we ought not discriminate between the suffering/pleasure of an AI and a human being.*

The combination of both the previous statements blocks the first suggestion mentioned: if one accepts the possibility of supersuffering, then all that should matter for the long-term development of AI is the issue of supersuffering and that of superpleasure.

Conclusion 2. *From an ethical point of view, the possibility for an AI to experience supersuffering takes precedence over the expected benefits that an AI will produce for mankind.*

The Unique Responsibility of Creating an AI

In order to show how the consequences of creating an AI are unlike the consequences of other acts that we collectively engage in, I will make explicit in what sense humanity would be responsible for supersuffering if it were to occur. The notion of responsibility that I have in mind is morally neutral, in the sense that it does not by itself imply blame or praise. Rather, it implies that one is a possible candidate for receiving blame or praise: if the outcome is negative, then a *response* is required from the agent in order to justify the negative outcome. If no proper response can be given, then the agent is blameworthy.

For example, imagine that a doctor decides to operate a patient who does not have a life-threatening condition. The operation is fairly safe, but still there is a very small probability that the patient will not survive it. Unfortunately, in this particular case the patient does indeed die, due to factors that were beyond the doctor's control. Here the doctor is responsible for the patient's death, in the sense that he would need to justify why performing the operation was the best decision. Doing so could take the form of comparing the prior probability of the patient's death with the increased comfort that a successful operation would have produced.

Premise 7. *If an agent (or a group of agents) knows that performing an act (or a set of acts) A might cause an outcome O, and the agent(s) also knows that there exists an alternative act (or a set of acts) A' such that performing A' will certainly not cause O, then the agent(s) will be responsible for O if A turns out to cause O.*

We already concluded that humanity might cause supersuffering. In order to invoke the above premise, we need to add the following trivial counterpart.

Premise 8. *If all of humanity does not attempt to create an AI, then the set of our acts will certainly not cause an AI to ever experience supersuffering.*

We now arrive at our third conclusion.

Conclusion 3. *If humanity creates an AI, then we will be responsible for all supersuffering it might endure.*

On the short term, and when considering a single agent, there are many outcomes for which one is responsible in the sense of Premise 7. This no longer holds if we consider all of humanity and extend our horizon into the far future. Given our limited knowledge of the world, and the almost infinite complexity of the causal chain that results from our actions, we are ignorant with respect to the long-term consequences of our actions on the well-being of humanity.

Premise 9. *For any set of acts A that humanity performs today, to the best of our knowledge, it is possible that A will cause extreme human suffering in the long run.*

Combining this premise with Premise 7 gives:

Conclusion 4. *There is a time t such that even if the current acts A performed by humanity will cause extreme human suffering after t, we will not be responsible for this.*

This conclusion rules out the second suggestion: the reason why we find it acceptable that our acts might have extremely negative consequences for humanity in the future, is that *this holds just as well for any alternative acts that we might perform*. All we can do is focus on outcomes in the near-future, and hope for the best in the long-term. The distinguishing feature of our attempt at creating an AI is that this is no longer true, for *there is an obvious alternative act* that will certainly not cause supersuffering, namely to stop doing any research on AI.

Moral Asymmetry

At this point we can draw the following worrisome conclusion.

Conclusion 5. *It is possible that by creating an AI, we will be responsible for the greatest suffering that our world has ever known.*

Still, an optimist might argue, completely analogous to this depressing conclusion, we could also be responsible for the greatest pleasure that our world has ever known. Hence the route for the third suggestion to defend our attempt at creating an AI is still open.

I invoke a moral principle that as far as I can tell is accepted by the majority of mankind, to counter the strict utilitarian calculus that is employed in this argument. Nevertheless, I submit that a strict utilitarian could reject the following, in which case the argument does not go through.

Premise 10 (Moral asymmetry)**.** *All other things being equal, the moral blameworthiness for being responsible for someone's suffering to an amount X is greater than the moral praiseworthiness for being responsible for someone's pleasure to an amount X. Further, the difference between the degree of blame and praise strictly increases with X.*

The above principle is similar in spirit to the medical principle "first do no harm", and is confirmed by the moral risk-aversion that is widespread in our behaviour. For example, assume you may press a button such that with probability 0.5 a random person's leg will be broken, and with probability 0.5 someone's broken leg will be healed. I think it goes without saying that it is immoral to press the button.

Or imagine that a reliable but very powerful person offers you the following bet: he flips a coin, and if it lands heads then your best friend will become extremely rich, but if it lands tails then he will make sure that your friend will remain poor for the rest of his life. Even if the odds are slightly changed in favour of becoming rich, I take it that almost everyone would find it immoral to accept the bet. Many more examples can be easily constructed to illustrate this point.

In line with the earlier comment regarding the need of a precise estimate for the probability that an AI will supersuffer, there is no prima facie reason to assume that the probabilities of supersuffering and that of superpleasure vary greatly.

Premise 11. *The probability that we will be responsible for creating an AI that will supersuffer is of the same order of magnitude as the probability that we will be responsible for creating an AI that will enjoy superpleasure.*

All of the above results in the following conclusion.

Conclusion 6. *If we attempt to create an AI, our expected blameworthiness is much higher than when we do not attempt to create an AI.*

The following is part of the very meaning of what it means to be blameworthy.

Premise 12. *Humanity should act so as to minimise our expected blameworthiness.*

Which brings me to the final, quite dramatic, conclusion.

Conclusion 7. *Humanity should not attempt to create an AI.*

Conclusion

I have sketched the contours of an argument that if successful, would cast a dark shadow over a field that for many of us holds a far greater promise of contributing to the good than to the bad. However I believe this result might turn out to have an unexpected consequence.

Note that I do not draw any conclusion regarding the attempts of *individual human beings* to create an AI. In fact it is entirely consistent for a person to accept Conclusion 7 and despite this continue to believe that he or she should attempt to create an AI (or at least contribute to such an attempt in a very small way). Let me explain.

A convincing case could be made that regardless of the strength of the above argument, there will always remain a substantial group of people who will continue to work on AI. Specifically, the less ethical an AI researcher is, the less he or she cares about ethical concerns involving AI, and therefore the more likely that this person will belong to that group. Therefore if one considers oneself as a person with an above average concern for ethical issues, then this argument would be turned upside down!

Concretely, if one assumes that the probability of humanity's long-term success at developing an AI is independent of the amount of people currently working on it, then the decision of any contemporary AI researcher to leave the field has no impact whatsoever on whether or not an AI is ever created. The only impact such a decision would have is that by leaving the field, the researcher no longer has any control over *how* humanity attempts to create an AI.

So under this assumption, any person who both accepts the above argument, and accepts the ethical severity of the creation of a supersuffering AI, is under a moral obligation to continue working in AI in such a manner that the probability of supersuffering ever being instantiated is minimised. In other words, rather than an argument against AI research, this argument would turn out to be a plea for giving priority to ethical concerns in AI research. Although I do find the above assumption plausible, I will not defend it here but instead conclude that either one should not be involved in attempting to create an AI, or one should give priority to preventing the creation of a supersuffering AI.

Acknowledgements

Sander Beckers was funded with a Fellowship from the Belgian American Educational Foundation (B.A.E.F.).

References

Bostrom, N., and Yudowsky, E. 2014. *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press. chapter The Ethics of Artificial Intelligence.