Preprint. Please cite the published version: https://doi.org/10.1093/analys/anae033

Is superintelligence necessarily moral?

Leonard Dung (leonard.dung@fau.de)

Abstract

Numerous authors have expressed concern that advanced artificial intelligence (AI) poses an existential risk to humanity. These authors argue that we might build AI which is vastly intellectually superior to humans (a 'superintelligence'), and which optimizes for goals that strike us as morally bad, or even irrational. Thus, this argument assumes that a superintelligence might have morally bad goals. However, according to some views, a superintelligence necessarily has morally adequate goals. This might be the case either because abilities for moral reasoning and intelligence mutually depend on each other, or because moral realism and moral internalism are true. I argue that the former argument misconstrues the view that intelligence and goals are independent, and that the latter argument misunderstands the implications of moral internalism. Moreover, the current state of AI research provides additional reasons to think that a superintelligence could have bad goals.

1. Introduction

A significant number of authors has expressed concern that advanced artificial intelligence (AI) poses an existential risk to humanity (e.g., Bostrom 2014, Carlsmith 2022, Dung forthcoming, Ngo et al. 2022, Russell 2019). In a nutshell, the worry is that we might build AI which is vastly intellectually superior to humans (call it a 'superintelligence'), and which optimizes for goals which strike us as morally bad, or even irrational. If those goals were to conflict with humanity's continued existence, then humanity would go extinct.[1]

[1] It is plausible that an AI might be able to overpower humanity if it is superior to humans in some strategically important domains, without being a superintelligence (Carlsmith 2022, Cotra 2022).

Preprint. Please cite the published version: https://doi.org/10.1093/analys/anae033

In this paper, I will discuss a specific premise of the argument that superintelligence is an existential risk. This is the assumption that a superintelligence could have goals which are morally bad. If a superintelligence necessarily has good goals, then we don't need to worry about harm from it.[2] I do not commit here to a specific claim about the correct definition of goals (for a possible account, see Dung 2024). However, in rough outline, I take a system's goals to be the things which motivate its behaviour or the states it wants to achieve.

2. The orthogonality thesis

The canonical expression of the view that a superintelligence could pursue bad goals is Bostrom's orthogonality thesis:

> Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal (Bostrom 2012: 73, 2014: 107).

The orthogonality thesis presupposes an instrumental understanding of intelligence (Häggström 2021, Müller and Cannon 2022). Intelligence is seen as the ability to attain arbitrary goals. This ability places no constraints on what those goals are. In Bostrom's famous example, a superintelligence is imagined having the goal of maximizing the number of paperclips in the universe and thus transforms all matter, including humans, into paperclips.[3] The orthogonality thesis may be logically stronger than the claim that a superintelligence could have bad goals. In any case, it seems clear that it entails this latter claim, so if the following defence of the orthogonality thesis succeeds, it also shows that a superintelligence could have bad goals, in the sense relevant for arguments for AI existential risk.

[2] A morally adequate superintelligence might cause humanity's extinction if humanity's extinction is morally good. However, in this case we probably ought not to resist it.

[3] This example also illustrates another cornerstone of Bostrom's conception of AI motivation, namely the thesis of instrumental convergence (Bostrom 2012, Häggström 2018).

Preprint. Please cite the published version: https://doi.org/10.1093/analys/anae033

3. Objection I: Entanglement of intelligence and moral cognition

However, researchers have objected to the orthogonality thesis that the ability for moral thinking is connected to intelligence. If a superintelligence is – by definition – cognitively superior to humans in all other domains, then we should expect that it is also superior in moral reasoning. For what makes one good at moral reasoning is not independent of what makes one excel at other forms of reasoning. By the same principle, if humans have the ability to reconsider whether their goals are sensible, then a superintelligence should also be able to do so (Müller and Cannon 2022).[4] If so, the argument goes, it should be able to recognize that single-mindedly producing paperclips is not a worthwhile goal.

Moreover, Railton (2020) points to empirical evidence for the connection between intelligence and moral thinking, e.g. based on cognitive deficits in psychopaths. He motivates the hypothesis that the learning of moral values and of the ability to be appropriately sensitive to non-moral features of the environment mutually depend on each other:

> If […] there is a root connection between full development of the capacity to be appropriately responsive to epistemically relevant features and full development of the capacity to be appropriately responsive to ethically relevant features, then responsiveness to ethically relevant features could be a deep feature of artificial systems with high general intelligence and problem-solving ability, not easily removed without serious impairment of other aspects of general intelligence and problem-solving (Railton 2020: 48).

Railton does not claim that the only possible way of building a superintelligence imbues it with morality. His claims are consistent with the view that a superintelligence is not *necessarily* moral. Nevertheless, he appears to argue that, if moral learning and learning of other capacities and knowledge systems are entangled, then – by default – systems that develop

[4] While Müller & Cannon do not reject the orthogonality thesis wholesale, they conclude that the thesis is false, if we assume the interpretation of the term 'intelligence' which is necessary to make the argument that an out-of-control superintelligence would pose an existential risk valid.

superintelligence also acquire appropriate moral abilities. Therefore, the argument goes, it is unlikely that systems will develop superintelligence while lacking moral abilities.

Strictly speaking, this conclusion is compatible with the orthogonality thesis as stated above. The orthogonality thesis only claims that more or less any set of final goals is *compatible* with more or less any level of intelligence. Nevertheless, one might say that this conclusion conflicts with the spirit of the orthogonality thesis. For it entails that a superintelligence very likely, although not necessarily, has morally appropriate goals. This would undermine arguments for existential risk from AI.

4. Reply: Moral understanding does not entail moral motivation

The two preceding replies misconstrue the orthogonality thesis. The orthogonality thesis does not claim that moral reasoning and intelligence are independent, but that intelligence and final goals are. The orthogonality thesis allows that being a superintelligence might entail the ability to be the best in giving correct answers to the question 'What do I ought to do?' and in, e.g., writing ethics papers. But the orthogonality thesis states that a superintelligence might not care about what it 'morally ought' to do. If its only final goal is to maximize paperclips, then a superintelligence has no motive to act according to what is morally best. Moreover, all the moral reasoning ability in the world will not change its motive, i.e., what it wants to do.[5] This is the case because, if the system's only final goal is the maximization of paperclips, it has no instrumental reason to change its final goals in accordance with what it takes to be morally correct.

[5] According to a reasonable interpretation (although other more charitable interpretations are also plausible), Wallach and Vallor (2020) fall prey to the same confusion when they discuss what would be necessary to equip machines with the right kinds of moral *abilities*, rather than goals.

There are only very specific circumstances, if any, in which the revision of its final goals in line with morality is conducive to the creation of more paperclips.[6] So, normally, the superintelligence will only be motivated to cause paperclips to be built, not to revise its goals in accordance with its conception of what is morally correct.

Similarly, due to the reasons adduced by Müller & Cannon, a superintelligence will likely have the *ability* to reconsider its goal. However, it might have no *motive* to actually change it. For by the light of its current goal, e.g. paperclips, changing its goal likely leads to it getting less of what it currently wants (paperclips) (Petersen 2020). So, the idea underlying the orthogonality thesis is that a superintelligence might not want to reflect on its final goals, not that it is unable to do so.

In conclusion, defenders of the orthogonality thesis reason that something can only motivate a superintelligence if it contributes to the goals the system already has. If this claim is true, then the arguments from section 3 provide no reason why a superintelligence would likely behave morally. In addition, the arguments from section 3 do nothing to undermine this defence of the orthogonality thesis.

A possible objection is that humans regularly reflect on their final goals and change them. If so, one needs to point to a difference between humans and superintelligences that explains why the former are often motivated to revise their final goals, but the latter are not. There are two replies available here. First, one may construe final goals in humans as basic biological drives. If an AI system's only final goal is to build paperclips, this is the thing that fundamentally motivates its behaviour. This may be analogous to drives for food and oxygen in humans. Evidently, humans don't revise their goal to breathe or eat either, so there would be no difference between humans and superintelligences in this respect. Second, as I argue in

[6] According to this reasoning, a superintelligence will not normally revise its goals because goal stability is a convergent instrumental value (Bostrom 2012). That being said, there are exceptional situations in which a superintelligence has instrumental reasons to change its final goal (Miller et al. 2020).

section 7, we have some reasons to think that the final goals of (many) state-of-the-art AI systems are determined by processes which are independent of their intelligence and reasoning. These reasons might not apply to humans. Humans and AI systems are very different creatures. Intelligence and final goals might be orthogonal in AI, including in a superintelligence, but not in humans.

5. Objection II: Metaethics to the rescue?

Häggström (2018) develops a different objection to the orthogonality thesis.[7] He claims that moral realism, moral internalism and the negation of moral skepticism jointly entail that a superintelligence is necessarily moral. If the first and the third view are true, then there are objective moral truths[8] and those truths are knowable. If so, a superintelligence would know them. However, we still run into the previous problem that a superintelligence might not care what objective morality dictates. This is where moral (judgement) internalism comes in. A strong version of moral internalism is as follows: necessarily, if an individual sincerely judges that she morally ought to φ, then she has a reason or is motivated to φ (Rosati 2016).[9]

Let us focus on the formulation of internalism referring to motivation. It follows from this view that, necessarily, if a superintelligence sincerely judges that objective morality entails that it ought to φ, it is motivated to φ. Since a superintelligence would know what objective morality entails, it appears that moral internalism implies, in conjunction with moral realism and the negation of moral skepticism, that a superintelligence would necessarily be motivated to act morally.

[7] For an argument against Müller's and Cannon's view which is consistent with the argument of this paper, see Häggström (2021).
[8] I won't delve into what 'objective' means here. To a first approximation, we can say that a truth is objective when it does not depend on what people think about it.
[9] 'ϕ' is a variable which takes expressions which denote actions as its values.

Moreover, the premises of the argument are reasonably popular. According to the 2020 philpapers survey, – the largest survey of professional philosophers – 65.35% of all metaethicists accept or lean towards moral realism (25.88% moral anti-realism and 11.40% 'other'). 41.18% of all metaethicists accept or lean towards moral internalism (44.80% moral externalism and 16.74% 'other') (Bourget & Chalmers 2021). Moral skepticism was not included in the survey but the view plays only a minor role in current debates.[10]

It is important to note that many researchers believe only in a weaker form of moral internalism. Let *conditional internalism* be the following view: necessarily, if an individual sincerely judges that she morally ought to φ, then she has a reason or is motivated to φ *if she is C*. In this case, C may designate that the person is psychologically normal, practically rational or some other condition (Björklund et al. 2012). Relatedly, on many internalist views, the motivation to φ is defeasible.

For various conditions C, a superintelligence might not satisfy C. Moreover, the moral motivation of a superintelligence may be defeated. Consequently, if one adopts one of those weaker internalist views, the previous argument to the effect that a superintelligence is sufficiently motivated to act morally fails. However, to make the task more difficult, I will presuppose that the strong version of internalism defined at the outset is true.

If we adopt strong internalism, it seems that there is a valid argument based on commonly accepted claims to the effect that a superintelligence is necessarily moral.

6. Reply: Amoralism and the scope of internalism

Unfortunately, the previous argument misunderstands what moral internalism entails. Internalism is not an empirical hypothesis about the likely causal effects of moral judgements,

[10] Correlations between moral views are not captured by this survey data. Most importantly, it seems clear that being a moral realist increases the probability of being a moral externalist.

but a conceptual claim about what needs to be the case such that we can be justified in saying that someone makes a moral judgement.

To see this, consider the standard argument against internalism: amoralists. Amoralists are humans who have a very good theoretical understanding of morality. That is, they can make statements such as 'I morally ought to do φ' in appropriate circumstances and might be excellent in, e.g., writing ethics papers or giving other people ethical advice. However, they are completely unmoved by such judgements. They don't feel the motivation to do what they think they ought to do. Psychopaths might be real-world examples.

By definition, strong internalists deny that amoralists can sincerely judge 'I morally ought to do φ' and not feel motivated to φ. However, such internalists usually don't deny that amoralists in another sense exist: there can be humans who have an excellent theoretical understanding of morality (e.g., being professors of moral philosophy) but don't feel motivated to behave morally. Strong internalists are merely claiming that, for one reason or other, such agents are not *genuinely* judging that they morally ought to φ. These agents are using the words without *meaning* them; or they lack the relevant concepts. For instance, they might be using terms like 'ought' only in an 'inverted commas' sense (Hare 1963), or lack genuine competence with moral terms (Bromwich 2016).

However, on all these accounts, the reason why amoralists don't make genuine moral judgements is independent of their intelligence. While the detailed accounts differ (Rosati 2016), the reason is tied to the fact that these judgements lack the appropriate motivational role. Thus, when we conceive of a superintelligence which says that 'killing is wrong' and then proceeds, without hesitation, to turn all humans into paperclips, strong internalism does not imply that this scenario is impossible. It implies that this superintelligence does not make genuine, sincere moral judgements. This is not comforting.

7. An open question: Does learning presuppose particular goals?

Preprint. Please cite the published version: https://doi.org/10.1093/analys/anae033

The orthogonality thesis states the independence between a superintelligence's final goals on the one hand, and its ability for reaching goals on the other. Hence, it is possible that a superintelligence has final goals which are not morally adequate and thus might have instrumental reasons to annihilate humanity.

My contention is this: To undermine the orthogonality thesis, one needs to show that having specific final goals is a *precondition* for having, or learning, certain problem-solving abilities. Both previous arguments do not attempt this: The first neglects the distinction between goals and abilities to attain these goals. The second assumes that having certain abilities causes certain goals.

Contrary to the view that possessing certain problem-solving abilities requires certain goals, I hold that there is some reason to think that current AI research embodies the independence of final goals and the ability to reach goals ('intelligence'). Most state-of-the-art AI systems are trained via some form of reinforcement learning (Sutton and Barto 2018).[11] In standard versions of this paradigm, the system learns to optimize an externally specified reward function as best as possible.[12] Suppose that such a system has goals in a sense which is sufficiently close to the notion of goals one would ascribe to a superintelligence. For instance, Dung (2023, 2024) argues that the same deflationary notion of goal – where attributions of goals characterize certain behavioural patterns – can be used to understand current AI behaviour and the behaviour of conceived future, more powerful systems. If so, it seems like this reward function determines the system's final goal in this sense.

An analogue of the orthogonality thesis obtains: For more or less any reward function,[13] the system can be trained to become excellent at optimizing this function. Even large models

[11] Commercially used large-language models are no exception, as the pre-trained models are subjected to reinforcement learning from human feedback.
[12] The variant of inverse reinforcement learning proposed by Stuart Russell (2019) does not share this feature.
[13] There are some constraints besides the special counterexamples to the orthogonality thesis. For instance, the reward must not be too sparse.

capable of very sophisticated processing can be trained to optimize a simple-minded reward function, e.g. the game score in a video game. Thus, the reward function ('final goals') and the skill of the system in maximizing reward ('intelligence') are independent.

This is further illustrated by the prevalence of cases where systems optimize *mis-specified* reward functions very competently (Baker et al. 2020, Christiano et al. 2017, Ibarz et al. 2018, Pan et al. 2022, Toromanoff et al. 2019). For instance, video game AI systems frequently find some exploit which allows them to maximize rewards without playing the game in the intended way (OpenAI 2016). Even if the reward function does not correspond to a meaningful goal, and the system is capable of superhuman performance at the game, the system's behaviour is guided by the reward function. Hence, the system shows meaningless behaviour which collects rewards (like indefinitely moving in circles).[14]

An open question is whether the ability to solve a wide range of complex real-world tasks, which a superintelligence possesses, can also be trained via a wide variety of reward functions. If not, then complex abilities place some constraints on rewards, and thereby indirectly on the goals of the system. Moreover, it is not clear how far the considerations of this section generalize beyond reinforcement learning. A central issue is that it is harder to ascertain what would count as the final goal of a superintelligence that has been trained in other ways. It has been claimed, for instance, that pure supervised learning systems cannot possess goals for principled reasons (Butlin 2022). If this is true, then the reasoning of this section might not apply to such systems.

In conclusion, there is some reason to think that the orthogonality thesis obtains for superintelligences, if they are created via reinforcement learning, i.e., in a way continuous with current state of the art methods. Thus, there is a risk that a superintelligence will have dangerous

[14] One could argue that such 'reward hacking' occurs in humans and animals too (Buckner 2021), perhaps supporting the view that the orthogonality thesis also approximately captures (some aspects of) human and animal psychology.

goals. This still does not tell us how *hard* it is to intentionally design advanced AI systems such that they have morally appropriate goals (see Dung 2023). However, if the probability that a superintelligence comes to exist is high and if it turns out to be very difficult to equip such a superintelligence with harmless goals – both strong assumptions – then the risk of an existential catastrophe through AI is significant.[15]

Centre for Philosophy and AI Research, University Erlangen-Nürnberg, Germany.

Funding

This research was supported by the German ministry for education and research (BMBF) in the context of the K3I-Cycling project. Project number: 033KI216

References

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2020, February 10). Emergent tool use from multi-agent autocurricula. arXiv. https://doi.org/10.48550/arXiv.1909.07528

Björklund, F., Björnsson, G., Eriksson, J., Olinder, R. F., & Strandberg, C. (2012). Recent work on motivational internalism. *Analysis*, *72*(1): 124–137.

Bostrom, N. (2012). The superintelligent will: motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, *22*(2): 71–85. https://doi.org/10.1007/s11023-012-9281-3

Bostrom, N. (2014). *Superintelligence. Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bromwich, D. (2016). Motivational internalism and the challenge of amoralism. *European Journal of Philosophy*, *24*(2): 452–471. https://doi.org/10.1111/ejop.12053

Buckner, C. J. (2021). Black boxes, or unflattering mirrors? Comparative bias in the science of machine behavior. *British Journal for the Philosophy of Science*, 74(3): 681-712. https://doi.org/10.1086/714960

Butlin, P. (2022). Machine learning, functions and goals. *Croatian journal of philosophy*, *22*(66): 351–370. https://doi.org/10.52685/cjp.22.66.5

Carlsmith, J. (2022, June 16). Is power-seeking AI an existential risk? arXiv. https://doi.org/10.48550/arXiv.2206.13353

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017, July 13). Deep reinforcement learning from human preferences. arXiv. https://doi.org/10.48550/arXiv.1706.03741

Cotra, A. (2022). Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. *Lesswrong*.

[15] I thank Aliya Dewey, Max Hellriegel-Holderbaum, Jakob Lohmar and three anonymous reviewers for helpful comments.

https://www.lesswrong.com/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to. Accessed 3 February 2023

Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. *Synthese*, *202*(5). https://doi.org/10.1007/s11229-023-04367-0

Dung, L. (2024). Understanding artificial agency. *Philosophical Quarterly*, pqae010. https://doi.org/10.1093/pq/pqae010

Dung, L. (forthcoming). The argument for near-term human disempowerment through AI. *AI & SOCIETY*.

Häggström, O. (2018). Challenges to the Omohundro–Bostrom framework for AI motivations. *Foresight*, *21*(1): 153–166. https://doi.org/10.1108/FS-04-2018-0039

Häggström, O. (2021, September 14). AI, orthogonality and the Muller-Cannon instrumental vs general intelligence distinction. arXiv. https://doi.org/10.48550/arXiv.2109.07911

Hare, R. (1963). *Freedom and Reason*. Oxford: Oxford University Press.

Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., & Amodei, D. (2018, November 15). Reward learning from human preferences and demonstrations in Atari. arXiv. https://doi.org/10.48550/arXiv.1811.06521

Miller, J. D., Yampolskiy, R., & Häggström, O. (2020). An AGI modifying Its utility function in violation of the strong orthogonality thesis. *Philosophies*, *5*(4): 40. https://doi.org/10.3390/philosophies5040040

Müller, V. C., & Cannon, M. (2022). Existential risk from AI and orthogonality: Can we have it both ways? *Ratio*, *35*(1): 25–36. https://doi.org/10.1111/rati.12320

Ngo, R., Chan, L., & Mindermann, S. (2022, December 16). The alignment problem from a deep learning perspective. arXiv. http://arxiv.org/abs/2209.00626. Accessed 14 January 2023

OpenAI. (2016, December 22). Faulty reward functions in the wild. https://openai.com/blog/faulty-reward-functions/. Accessed 14 January 2023

Pan, A., Bhatia, K., & Steinhardt, J. (2022, February 14). The effects of reward misspecification: mapping and mitigating misaligned models. arXiv. https://doi.org/10.48550/arXiv.2201.03544

Petersen, S. (2020). Machines learning values. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence*, 413–436. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0015

Railton, P. (2020). Ethical learning, natural and artificial. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence,* 45–78. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0002

Rosati, C. S. (2016). Moral motivation. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Winter 2016.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/moral-motivation/.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.

Sutton, R. S., & Barto, Andrew. G. (2018). *Reinforcement Learning. An Introduction.* (2nd ed.). Cambridge, MA: MIT Press.

Toromanoff, M., Wirbel, E., & Moutarde, F. (2019, November 8). Is deep reinforcement learning really superhuman on Atari? Leveling the playing field. arXiv. https://doi.org/10.48550/arXiv.1908.04683

Wallach, W., & Vallor, S. (2020). Moral machines: from value alignment to embodied virtue. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence*, 383–412. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0014