Application: Robotics, Cognition, Intelligence (TUM)

Is unregulated AI research a good idea?

Jonas Strabel · May 31, 2026

The scenarios once mostly restricted to science fiction movies have now moved into academia because of the unprecedented increase in Artificial Intelligence (AI)'s capabilities. While AI was laughed at for missing basic capabilities just a few years ago, frontier models are now even evaluated for whether they would sabotage AI safety research[1], or are not released to the public because they could potentially allow actors with malicious intent to cause significant harm[2]. In such times, questioning the continued development of AI is more urgent than ever.

1. Motivating the question

It is quite obvious that AI will have vast consequences for humanity. On the one hand, significant positive ones, such as advances in research that can lead to faster discoveries of cures and treatments. On the other hand, the outcomes are potentially extremely negative, going so far as the complete annihilation of humanity[3]. This creates a moral dilemma. In deontological / Kantian ethics, bad actions are often not allowed even if the result would be good; in these frameworks, what matters is not so much the outcome but the action itself[4]. In Rossian pluralism, there are multiple prima-facie duties to honour that have to be weighed against each other[5, ch. 2]. As a rule of thumb, non-maleficence outranks beneficence[5, ch. 2], so risking harm by developing AI is not justified even if it can do good. However, individual judgement in the situation is still required. Consequentialism (utilitarianism) demands that we bring the world into a state of greater good[6]. This means we have to innovate if doing so can save lives, but it also means we have to stop innovating if it causes harm. For the purpose of the following comparison, I will focus on the latter two frameworks, because they focus our attention on the potential outcomes. To resolve this dilemma, let us first evaluate our options as humanity. We need to decide whether to continue development, and if so, how. Let us call that regulation. Obviously, regulation is a very complex construct that can mean very different things, and the outcomes of various regulations are not always predictable. It is conceivable, for example, that forbidding certain kinds of development leads researchers to different problems that 'by accident' accelerate progress in AI (it would not be the first time such a thing has happened). Usually, however, taking resources away from a specific research area slows development in it: fewer people, less budget, fewer results. For the purpose of this paper I will consider exactly this aspect. Regulating shall mean slowing down research by providing it with fewer resources (similar to KatjaGrace's essay)[7]. Accelerating it then means the opposite: allocating more funding and more people to the goal. As humanity, we then have four options regarding regulation: (1) none, keep things as they are; (2) complete regulation, stop it; (3) moderate regulation, slow it down; or (4) anti-regulation, accelerate it[1].

[1]Note that (1) and (4) are similar, because no regulation will probably lead to acceleration anyway due to economic interest. (4) is nevertheless distinct, because humanity could decide (through governments) to invest even more and therefore accelerate it to a greater extent.

2. Game theory

Bostrom, Douglas, and Sandberg argued that what "unregulated" means is simply that the most optimistic actor acts: if many actors can each independently decide to build, then the most 'reckless' one will move first[8, pp. 350–351]. To regulate, then, means to move not at the speed of the most reckless, but at that of the majority. It is often said that "if we don't build it, they will". However, that is a false dichotomy, as it is not as though communication with the other side is impossible. Continuing this rapid innovation without any regulation could lead to unexpected outcomes, just as hardly anyone predicted the explosion in the linguistic capability of language models that has occurred in recent years. These outcomes could also be lethal. A 2023 survey of 2,778 AI researchers gave a median probability of 5% to extinction-like outcomes from AI[3]. Notably, this survey included researchers not just from the US, but also from China[3]. When both parties realise that racing towards the goal is a loss for both of them, communication might become possible, because it is in both of their interests.

To summarise, there are multiple games one could play here[7, inspired by].

a) *The arms race (prisoner's dilemma):* This game presupposes that there are two or more participants, each inevitably racing to build an AI that will make them vastly superior to their opponents. If one builds it and the others do not, the one who builds it gains an extreme advantage. If both build it, the scores are worse than the baseline (nobody building it), because both invest with no advantage.

b) *The suicide race:* Here, building such an AI leads to a fatal outcome for the whole of humanity. This means that if either party, or both, builds such an AI, humanity as a whole loses. The only good outcome is nobody building it.

c) *The safety-or-suicide race:* In this scenario, AI is detrimental for the party that does not build it if the other party builds it without restraint. But it is less detrimental, indeed even best, for both parties (because of the many advantages of the technology) if they both build it collaboratively and slowly. Not building it is the zero baseline.

It seems clear that a) is the wrong game, because AI advancements could lead to unexpected lethal outcomes, so even if one party were faster, that would not mean it wins, rather, it loses too. This goes in the opposite direction as well: if AI turns out not to harm us but to benefit us, the outcome is much better than the baseline for both parties. This leads to b). In that case, however, the only sensible thing would be not to build at all. This in turn ignores the possibility of a third path: c). Here, the gains (which, according to the ethics above, we should strive for) would still be won (even if later) but without the detrimental result, or at least with a reduced one.

3. Conclusion

It seems clear that a) is the wrong model for the situation. b) is the worst-case model, which should only be considered if no collaborative regulation is possible. This leaves c) as the clear winner. Within c), 'both build, but safely' seems to be the best path for humanity as a whole. So no, unregulated research is not a good idea. There is, however, an important caveat regarding regulation. In the sense I use the word in this paper, namely, developing more safely and more

slowly, the conclusion is likely true. But, as noted earlier, regulation can also have adverse effects[9]. What if we, as humanity, decide to regulate, but in doing so trigger even faster innovation with worse outcomes? That possibility cannot be excluded. There is therefore a strong need for innovation in regulation that actually leads to safer rather than less safe development[10, pp. 92–95]. But it is clear that regulation of some kind is the better path.

References

[1] Robert Kirk et al. *Evaluating Whether AI Models Would Sabotage AI Safety Research*. Apr. 27, 2026. DOI: 10.48550/arXiv.2604.24618. arXiv: 2604.24618 [cs.AI]. URL: http://arxiv.org/abs/2604.24618 (visited on 05/31/2026). Pre-published.

[2] *Project Glasswing: Securing Critical Software for the AI Era*. URL: https://www.anthropic.com/glasswing (visited on 05/31/2026).

[3] Katja Grace et al. "Thousands of AI Authors on the Future of AI". In: *Journal of Artificial Intelligence Research* 84 (Oct. 6, 2025). ISSN: 1076-9757. DOI: 10.1613/jair.1.19087. arXiv: 2401.02843 [cs.CY]. URL: http://arxiv.org/abs/2401.02843 (visited on 05/31/2026).

[4] Larry Alexander and Michael Moore. "Deontological Ethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2025. Metaphysics Research Lab, Stanford University, 2025. URL: https://plato.stanford.edu/archives/win2025/entries/ethics-deontological/ (visited on 05/31/2026).

[5] W. D. Ross. "What Makes Right Acts Right?" In: *The Right and the Good. Some Problems in Ethics*. Ed. by William David Ross. Clarendon Press, 1930.

[6] Walter Sinnott-Armstrong. "Consequentialism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2023. Metaphysics Research Lab, Stanford University, 2023. URL: https://plato.stanford.edu/archives/win2023/entries/consequentialism/ (visited on 05/31/2026).

[7] KatjaGrace. *Let's Think about Slowing down AI*. LessWrong. Dec. 22, 2022. URL: https://www.lesswrong.com/posts/uFNgRumrDTpBfQGrs/let-s-think-about-slowing-down-ai (visited on 05/31/2026).

[8] Nick Bostrom, Thomas Douglas, and Anders Sandberg. "The Unilateralist's Curse and the Case for a Principle of Conformity". In: *Social Epistemology* 30.4 (July 3, 2016), pp. 350–371. ISSN: 0269-1728, 1464-5297. DOI: 10.1080/02691728.2015.1108373. URL: http://www.tandfonline.com/doi/full/10.1080/02691728.2015.1108373 (visited on 05/31/2026).

[9] 1a3orn. *Ways I Expect AI Regulation To Increase Extinction Risk*. LessWrong. July 4, 2023. URL: https://www.lesswrong.com/posts/6untaSPpsocmkS7Z3/ways-i-expect-ai-regulation-to-increase-extinction-risk (visited on 05/31/2026).

[10] Brian Judge, Mark Nitzberg, and Stuart Russell. "When Code Isn't Law: Rethinking Regulation for Artificial Intelligence". In: *Policy and Society* 44.1 (Jan. 4, 2025), pp. 85–97. ISSN: 1449-4035. DOI: 10.1093/polsoc/puae020. URL: https://doi.org/10.1093/polsoc/puae020 (visited on 05/31/2026).

Declaration of Authorship

I hereby declare that I have written the present essay titled *Is unregulated AI research a good idea?* independently and have used no sources or aids other than those indicated. All parts of the work taken verbatim or in substance from published or unpublished writings have been identified as such.

Saulheim, May 31, 2026 _____

Location, Date Jonas Strabel