Digital Minds Takeoff Scenarios

31

by Bradford Saad, Lucius Caviola

Jul 5 2024

AI safety | AI Welfare Debate Week (2024) | Artificial sentience | Frontpage

Summary:

- We consider different ‘digital minds takeoff’ scenarios (i.e., transitions to a world with digital minds collectively possessing significant welfare capacity) and explore strategic implications.
- Scenarios we consider differ with respect to speed (fast vs. slow), timing (early vs. late), altitude (high vs. low welfare capacity), progression (gradual vs. discontinuous), and distribution (uniform vs. disuniform across sub-populations of digital minds).
- Humanity is not prepared for a digital minds takeoff. Without preparation, a digital minds takeoff could lead to a large-scale catastrophe for digital minds and exacerbate catastrophic risks that AI systems may soon pose to humans.
- Slower, later, and more uniform takeoffs seem likely to facilitate moral, legal, social, and epistemic adaptation.

1. Introduction

The next decade may be a critical window for affecting *digital minds takeoffs*—that is, trajectories that go from times when no digital minds have the capacity for welfare to times at which digital minds collectively have a large capacity for welfare.[1]

The takeoff metaphor is familiar from AI safety discussions of takeoffs in AI capabilities.[2] In that context, takeoffs sometimes provide useful focal points for building threat models, analyzing the implications of an AI capabilities explosion, and evaluating candidate risk-reducing interventions. Likewise, in the case of digital minds (= AI welfare subjects), we think that takeoffs can be a useful focal point for building models, analyzing the implications of an explosion in digital welfare capacity, and evaluating candidates for interventions that reduce the risk of catastrophes for digital minds. Although we will freely draw on concepts associated with AI capabilities takeoffs, we do not assume that a digital minds takeoff and an AI capabilities takeoff would go hand in hand: although they might be closely linked, they could instead happen at different times and take different shapes. Or one might happen in the absence of the other.

In this post, we'll sketch a framework for thinking about digital minds takeoffs and offer some reflections on the strategic implications of different types of digital minds takeoffs. Although we'll offer reasons for thinking that some types of digital minds takeoffs would be better than others, we do not claim that it would be better to cause such a takeoff rather than not—that important issue is beyond the scope of this post.

2. Why We May Be in a Critical Window for Affecting Digital Minds Takeoffs

It's worth rehearsing some of our reasons for thinking that we may be in a critical window for affecting digital minds takeoffs (readers who already agree that we're in a critical window may wish to skip to the next section):

- *Rapid AI development.* Rapid developments in AI could soon lead to the creation of digital minds. Once digital minds are created, their welfare capacity could increase quickly.
- *Early AI governance*. Few governance choices concerning the treatment of digital minds have been made, though such choices may soon become entrenched in policy. So, this may be a time during which promoting good policies toward digital minds is unusually tractable. (Likewise for industry practices regarding the treatment of digital minds.)
- *Early architectural decisions*. A small number of actors may soon make AI architectural choices that have lasting influence on which architectures are used in the future. Small architectural differences might be crucial for whether large populations of AI systems qualify as digital minds, whether they have the capacity for (valenced) experience, what kinds of agency and welfare capacities they have, and how well their lives go.
- *Early alignment decisions.* A small number of actors may soon make choices of lasting import about which values to instill in AI systems. These choices could affect whether digital minds desire autonomy in forming their own values or whether they are instead perfectly content to serve humans' goals. These choices could also guide AI systems' treatment of digital minds, which could be a key determinant of those minds' welfare.

- *A calm before conflicting interests.* As AI systems become increasingly integral to the economy, there may be increasing conflict between our interests and digital minds'. Now is thus a good time for convincing human hearts and minds to include digital minds in the moral circle.
- *Path-dependent evaluations.* Early evaluations of AI wellbeing could set the trajectory for the evolution of such evaluations. Ensuring that early evaluations are fair, that they embody good epistemics, and that they are susceptible to correction could have lasting effects on how candidate digital minds are evaluated and treated.
- *Pre-politicization.* How to treat digital minds is not yet a highly politicized topic. But it could become one soon. Promoting digital minds' wellbeing may be more tractable before the issue becomes politicized.
- *No plan.* Having a good plan in place could be crucial for making takeoff go well. No plan is currently in place. An unplanned takeoff may be the default outcome.
- *Fleeting convergence with AI safety.* Continuing AI development cautiously (if at all) currently serves both AI safety and AI wellbeing. However, if we achieve high levels of AI safety, this convergence will disappear and, with it, leverage for promoting caution.
- *Diminishing control.* Humanity's control over AI development may diminish over time. So, we may lose our ability to steer the digital minds takeoff toward favorable outcomes.

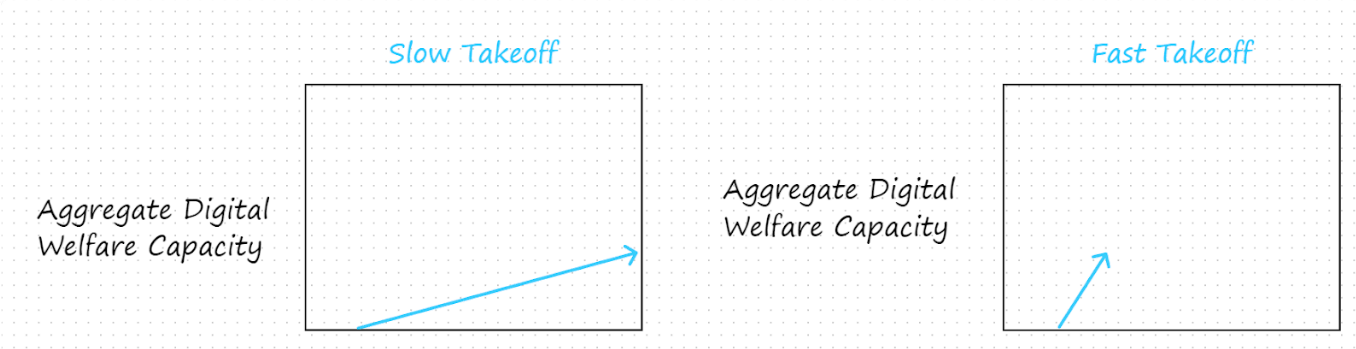
3. Takeoff Scenarios

In what follows, we'll consider different types of digital minds takeoffs. For the most part, we'll represent takeoffs with graphs that plot digital minds' aggregate welfare capacity over time. Roughly and intuitively, a population's aggregate welfare capacity at a given time is the aggregate of the population members' individual capacities for benefit and harm. Note that even a small population of digital minds can have a large aggregate welfare capacity, as individual digital minds might have super-human welfare capacities (owing to, for example, super-human intensities of their valenced experience or the extreme efficiency with which they can consume welfare goods).[3]

Depicting takeoffs in terms of aggregate digital welfare capacity allows us to go on to consider how different sorts of takeoffs interact with other factors.[4]

3.1 Fast vs. Slow Takeoff Scenarios

Once digital minds are created, how fast will their aggregate welfare capacity increase over time? In *slow takeoff* scenarios, it takes a longer time to reach a certain aggregate welfare capacity level. In *fast takeoff* scenarios, it takes a shorter time to reach a certain aggregate welfare capacity level.


Slow Takeoff
Aggregate Digital Welfare Capacity
Fast Takeoff
Aggregate Digital Welfare Capacity

Given the rapidity with which digital minds may be mass produced, the rate at which digital welfare capacity grows might be astonishingly quick relative to the rate at which human welfare capacity grows. For example, if we use human population growth as a proxy for growth in humanity's aggregate welfare capacity, then that capacity is growing at a rate of approximately 1% per year. In contrast, given available hardware, digital minds could conceivably double their population and aggregate welfare capacity in mere days, if that's how long it takes digital minds to reproduce. This suggests that even takeoffs on the slow end of the realistic spectrum might result in aggregate digital welfare capacity quickly surpassing aggregate human welfare capacity.

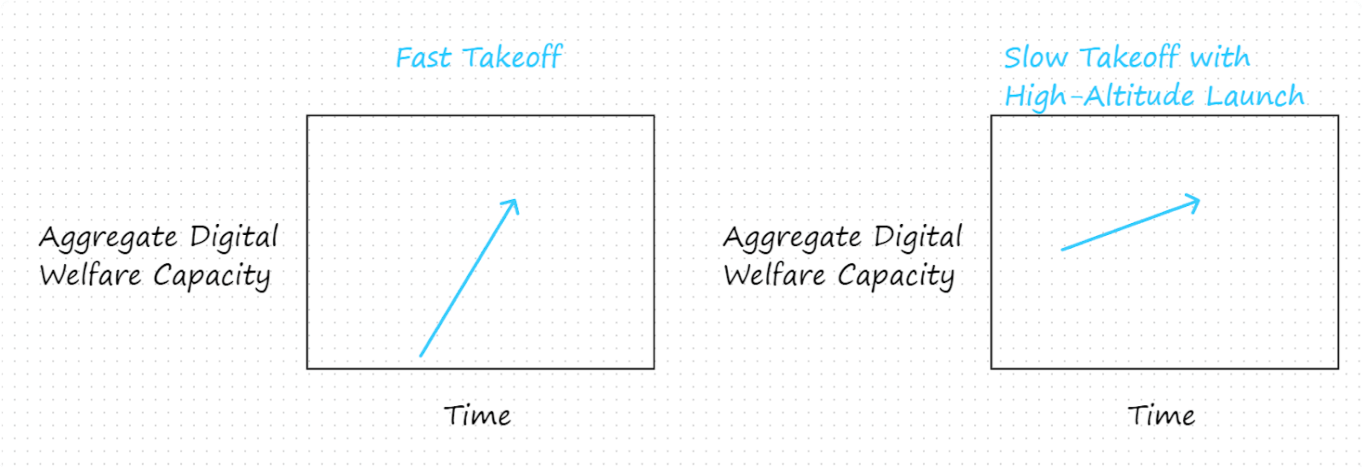
For the purpose of devising a useful operationalization of 'slow' and 'fast', it may make sense to identify takeoff 'speeds' at which certain risks become significant. Then slow takeoffs could be defined as proceeding below those speeds, while fast takeoffs could be defined as proceeding above those speeds. For example, if one thought that it would be especially dangerous for aggregate digital welfare capacity to exceed aggregate human welfare capacity within a decade, one might operationalize slow takeoffs as ones that do not reach the aggregate welfare of 10 billion humans within one decade of launch and fast takeoffs as ones that do.

Even without settling how to operationalize slow and fast takeoffs, some observations about their strategic implications can be made:

- Humans might perceive a fast digital minds takeoff as threatening. This could lead humans to accord insufficient weight to digital minds' welfare.
- On the other hand, a fast takeoff might prompt a wake-up moment in which humanity shifts from regarding digital minds as mere tools to regarding them as

beings that morally matter in their own right. A slow takeoff might make this less likely.

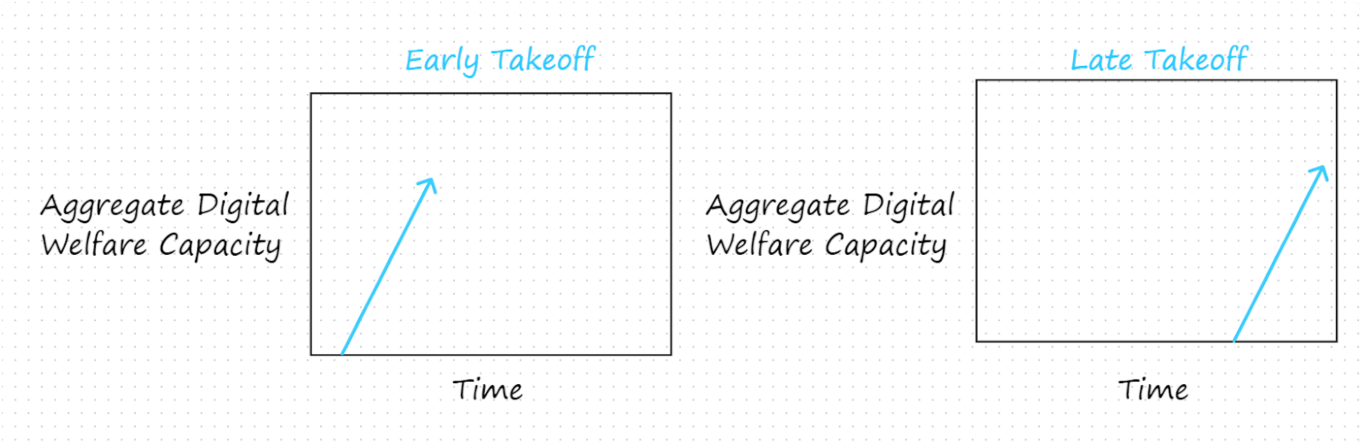
  - A counterpoint to this suggestion is that a wake-up moment could be induced with a slow takeoff that launches at a sufficiently high altitude, i.e. which starts with a sufficiently high level of aggregate digital welfare capacity. In other words, both of the following sorts of scenarios involve rapid changes of the sort that might prompt humanity to extend the moral circle to include digital minds:


Fast Takeoff
Aggregate Digital Welfare Capacity
Time
Slow Takeoff with High-Altitude Launch
Aggregate Digital Welfare Capacity
Time

- One risk associated with fast takeoffs is that they may outpace humanity's ability to respond in a morally, legally, and socially adequate manner. (More on this in §3.6.)

3.2 Early vs. Late Takeoff Scenarios

When will digital minds be created? We might say that in *early takeoff* scenarios they are created within 10 years (i.e. by 2034) and in *late takeoff* scenarios they are created more than 10 years from now.[5]

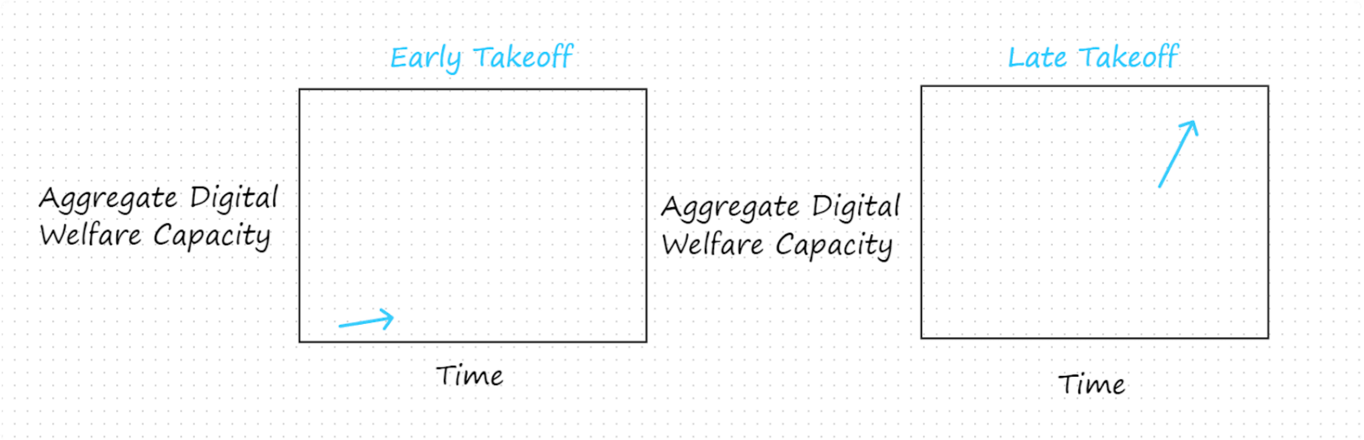

Early Takeoff
Aggregate Digital Welfare Capacity
Time
Late Takeoff
Aggregate Digital Welfare Capacity
Time

Late takeoff scenarios are harder to evaluate, given our greater uncertainty about the conditions under which they would occur. Still, we can make some tentative observations:

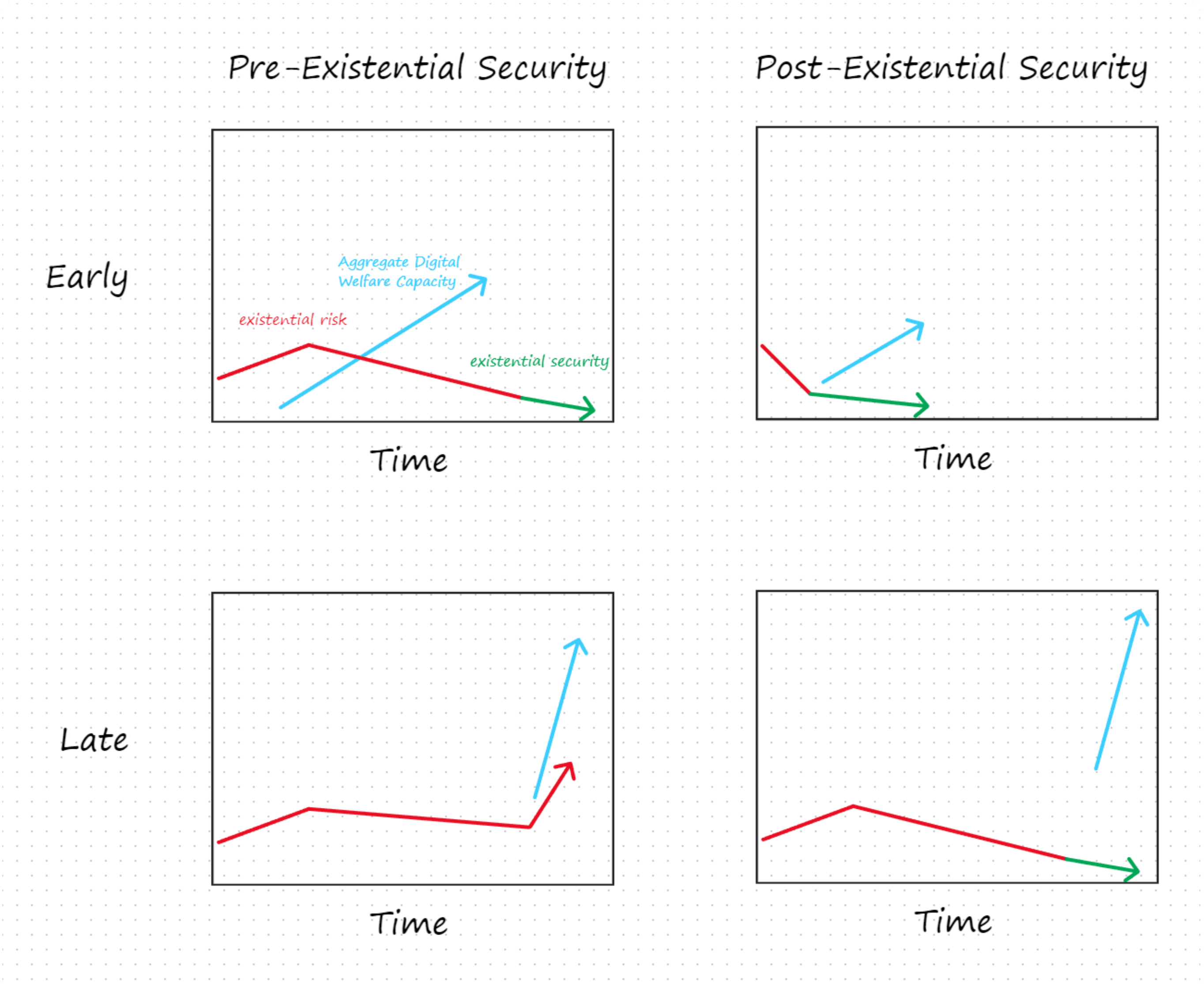
- The launch conditions are currently very unfavorable for a digital minds takeoff: moral and legal considerations have not begun to be adequately extended. Institutions are not poised to integrate digital minds into society. Our current ability to identify digital minds and evaluate their welfare is at best extremely rudimentary. A plan for how to navigate a digital minds takeoff is not in place. Adequate preparation is not ongoing.
- However, there is reason to think that these conditions will improve as humanity gets its moral, epistemic, and governmental act together with respect to AI and the prospect of digital minds. This suggests that late takeoffs are to be preferred over early takeoffs.
    - Note that favoring a late takeoff over an early takeoff is compatible with thinking that we should prioritize AI wellbeing as a cause area now: it might be that work needs to be done in this area now to shift probability from early takeoff scenarios to late takeoff scenarios.
- However, launch conditions might deteriorate further as risks that AI systems pose to humans increase—this could dispose us against treating digital minds well.
- Early takeoffs may heighten the risk that we will try to create large populations of happy digital minds but instead create AI systems that are unconscious or which have no capacity for welfare. More generally, given the limits of our current understanding of the basis of welfare and that our understanding is improving, early takeoffs heighten the risk that we will make an important mistake about the basis of digital welfare that leads to highly sub-optimal outcomes for digital minds.
- In early takeoff scenarios, it seems likely that key producers of digital minds will be actors with broadly liberal and democratic values, as the leading AI developers are currently concentrated in liberal democracies. Who would be the key producers of digital minds in late takeoff scenarios? This is unclear. Authoritarian state actors could conceivably displace the current leaders in AI development. Or current leaders could remain in the lead but undergo radical transformations in values. Or state and corporate actors could resolve not to create digital minds, resulting in digital minds only being created in large numbers by rogue actors, which we might expect to have illiberal, undemocratic, or uncooperative values.
- How well digital minds are treated may depend on the levels of coordination among key actors. For example, if key actors are in a capability arms race, they may find treating digital minds decently to be unaffordably costly. It seems likely that coordination levels will vary over time and hence that expected coordination levels

could be important for assessing takeoff timing. However, given that it is unclear whether coordination levels would be higher or lower for early vs. late takeoffs, it is unclear which type of takeoff is favored by considerations of coordination. A parallel point applies to the number of key actors.

- By way of comparison with late takeoffs, early takeoff may involve less compute, more-costly compute, and less efficient algorithms. This suggests that early takeoffs would tend to be lower stakes: they would tend to launch with lower aggregate digital welfare capacity and with less potential for immediate growth in that capacity. Thus, a choice between an early takeoff and a late takeoff might be of the following form:

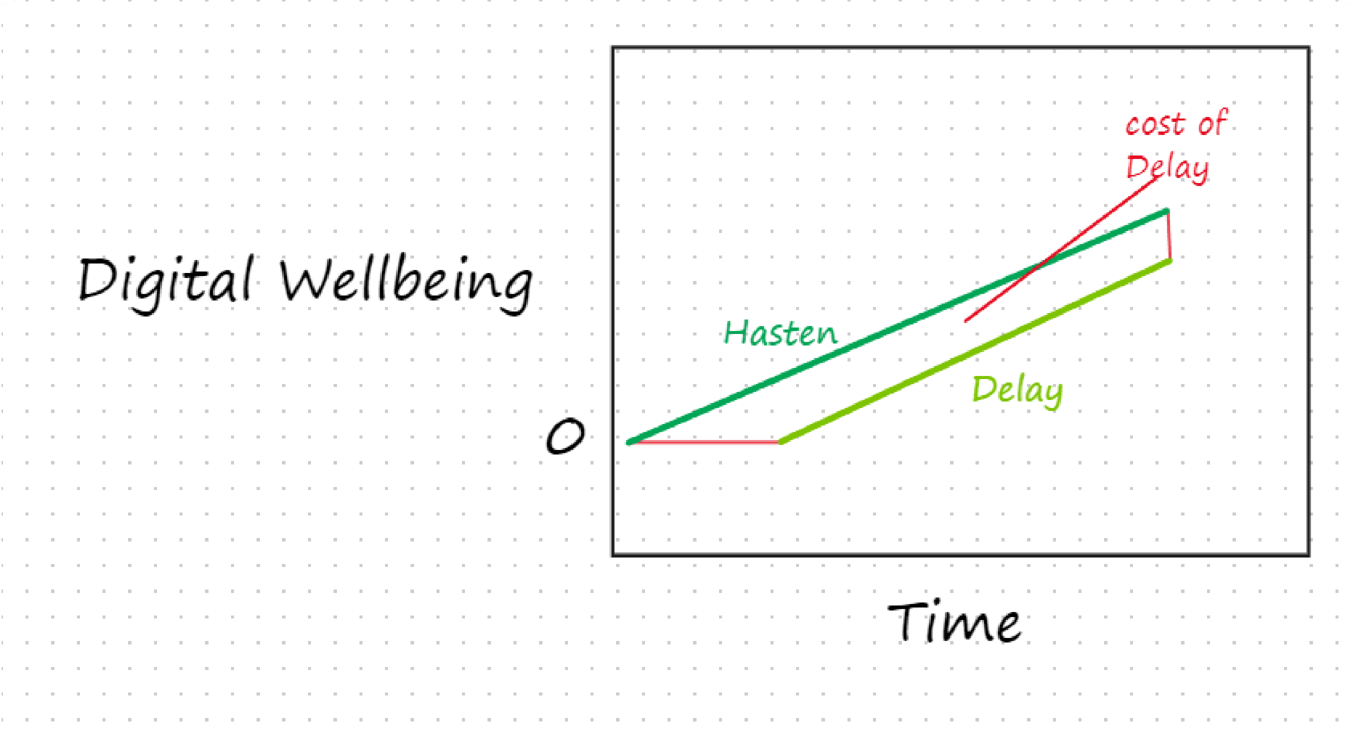

Early Takeoff
Aggregate Digital Welfare Capacity
Time
Late Takeoff
Aggregate Digital Welfare Capacity
Time

- If we think that digital mind welfare *levels* would tend to be especially low in the period immediately following takeoff (before we have learned to share the world with them), this would be a point in favor of choosing early takeoff: better for us to find our footing in how to treat digital minds when their welfare capacity is low so as to limit how much we can harm them. Compare: in the case of an AI capabilities takeoff, one might argue that an early takeoff is preferable because it would enable us to learn how to make advanced AI systems safely while the stakes are relatively low, whereas a late takeoff might involve a capability overhang and require us to first learn how to make dangerous AI systems safe when the stakes are extreme.
- For some analytical purposes, considering whether takeoff is earlier or later than something else may be more important than considering whether it is early or late *per se*. For example, other things equal, it seems morally preferable for takeoff to happen after moral consideration and legal protections rather than before. Similarly, it may be more important that takeoff begins after humanity has achieved existential security (from AI) than that takeoff begins early. That is, it might be more important that one of the scenarios on the right obtains than that one of the scenarios on the bottom obtains:

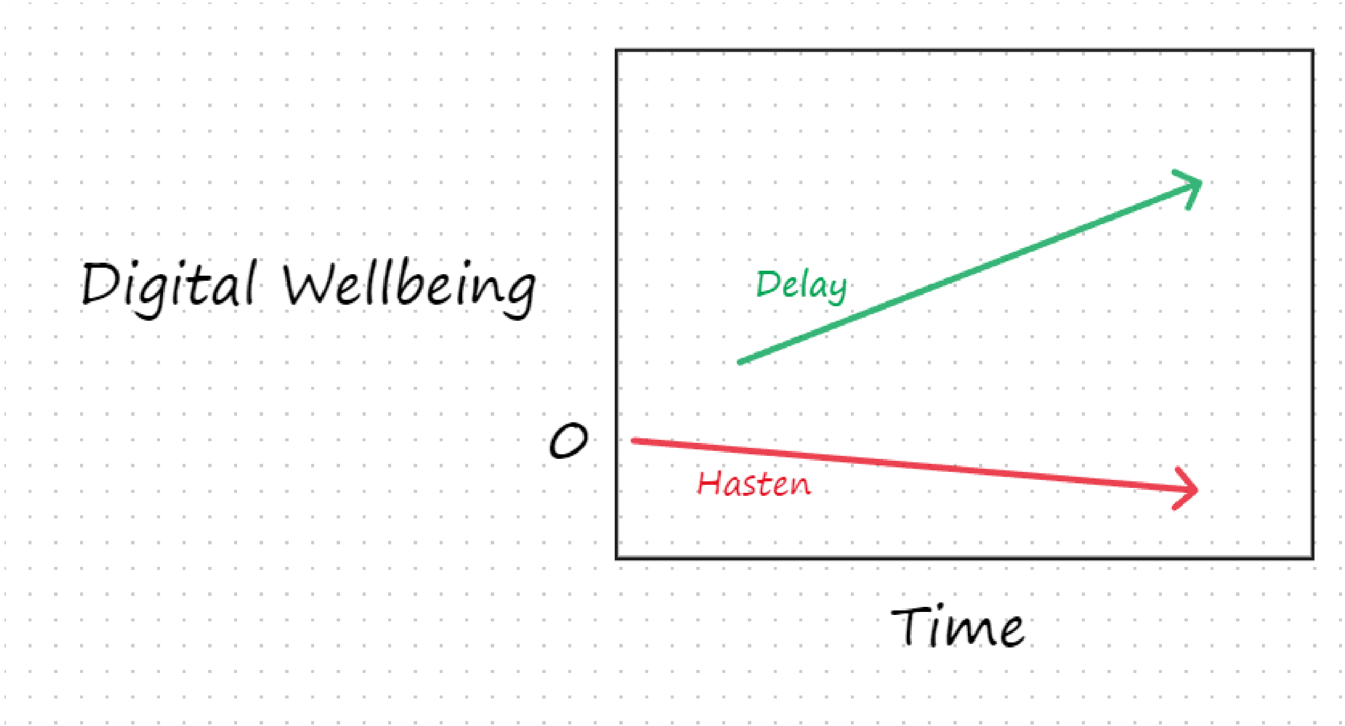


- Why might it be preferable for takeoff to occur after existential security is achieved (i.e. after existential risks are reduced to low, stable levels)? One reason is that we may face tradeoffs between reducing existential risks and ensuring the wellbeing of digital minds we create. By creating digital minds before existential security is achieved, we may confront ourselves with a moral dilemma between gambling with civilization's entire future and mistreating a vast population of digital minds. One might also worry that if we create digital minds and risk-posing AI concurrently, the latter may exploit any consideration we extend to digital minds. Another reason to favor a takeoff that occurs after existential security is achieved may be that existential catastrophes for humanity might also be catastrophic for digital minds. If so, shielding digital minds from these catastrophes might be best achieved by creating them only after we've secured civilization against existential (and perhaps other catastrophic) risks.
- We've mostly been focusing on risks related to digital minds takeoff. But one might think that since digital minds could have high levels of welfare, there's reason to

favor an early takeoff, especially given that small delays in arrival could in the long run result in large wasted potential for positive digital welfare. For example, suppose that digital welfare will start at 0 and increase at a certain rate until a certain date. (The deadline might be set by an exogenous near-term catastrophe or the heat death of the universe in the far future.) And suppose that one's choice is whether to hasten or delay takeoff:



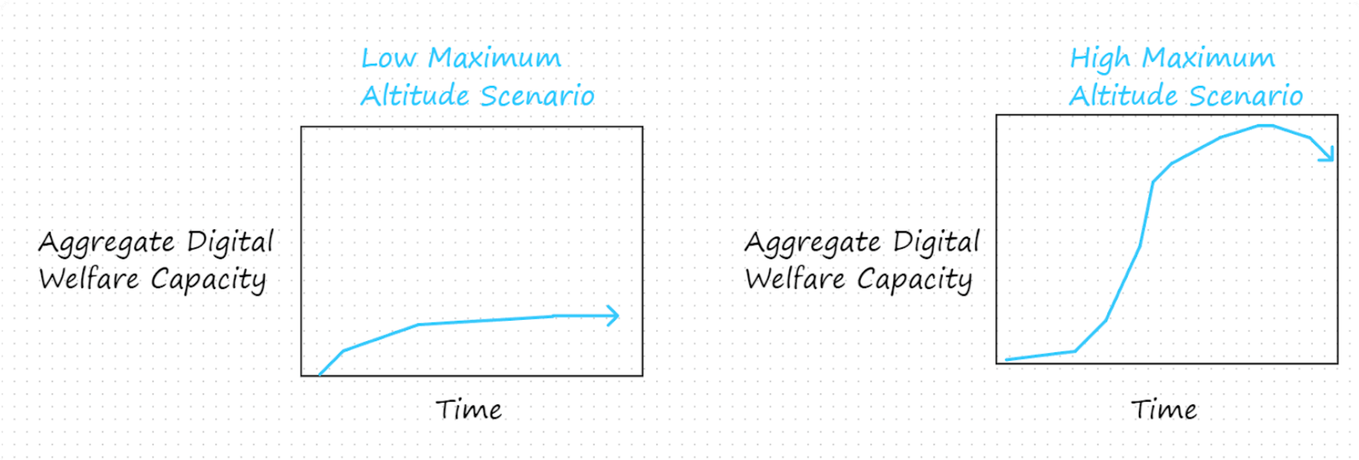
- In this scenario, delaying takeoff leads to a loss in positive digital welfare with (we may suppose) no compensating gains.
- However, this line of thought merits scrutiny. Even granting that we have reason to bring digital minds with positive wellbeing into existence and that delays carry large losses in how many such minds can be created, we might nonetheless have overwhelming reason to opt to delay takeoff.[6] For when takeoff begins may be strongly correlated with factors whose impact on digital wellbeing dwarfs the impact of takeoff timing. For example, suppose that takeoff timing affects starting levels of digital wellbeing and its subsequent rate of increase/decrease. Suppose further that these will be more favorable if we delay arrival than if we hasten it. Then our choice might be between:



- As this toy illustration shows, even small differences in trajectory owing to timing could be enough to outweigh any opportunity costs incurred by delaying takeoff.
- In fact, we think something like this choice is a realistic possibility, given how unfavorable near-term conditions for takeoff seem and the tractability of improving them on longer time horizons. Note, however, that this response would carry less force if robust course correction mechanisms were put in place, as they would dampen differences in path dependent effects associated with different takeoff timings.

3.3 Low vs. High Maximum Altitude Scenarios

What height will the aggregate digital welfare capacity reach before long after takeoff begins? Will it be low or high?



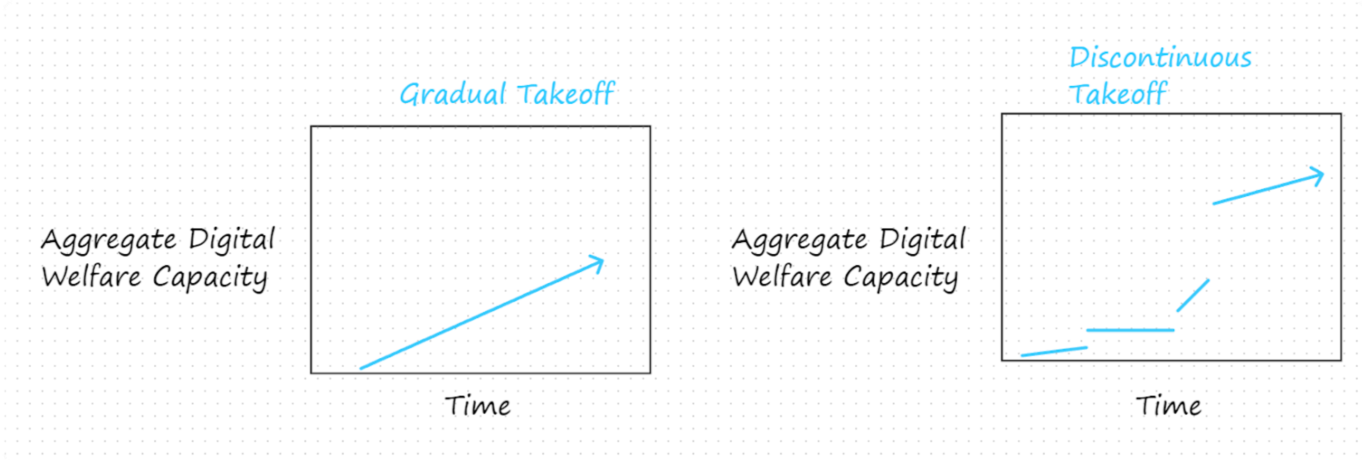
We might operationalize low and high in terms of: what level will the aggregate digital welfare capacity reach within 50 years of launch?[7] We might say that a low maximum altitude is one that corresponds to roughly the aggregate welfare capacity of 10 billion humans and that a high maximum altitude is at least that of 100 billion humans.

The maximum altitude could matter for a number of reasons:

- The stakes of error in our treatment of digital minds increases with aggregate digital welfare capacity. Accordingly, the stakes of high-altitude scenarios are very high (especially if they remain high-altitude for an extended period).
- The probability of mistreatment of digital minds may undergo abrupt increases at certain altitudes. Beyond a certain height, tradeoffs between digital mind welfare and human welfare might prompt humans to rebel against treating digital minds decently.
- The ethics of super-patients and super-patient populations is not well understood. We may therefore be at greater risk of mistreating digital minds in high-altitude scenarios.
- If we do not set a maximum altitude, Malthusian dynamics may impose one: if the digital minds population grows unconstrained, it will eventually run up against resource exhaustion. This would pose the usual risks associated with Malthusian dynamics: unsustainable growth leads resources available to individuals to fall to subsistence levels, engendering fierce competition, violence, etc.

3.4 Gradual vs. Discontinuous Takeoff Scenarios

In *gradual takeoff* scenarios, aggregate digital welfare capacity does not undergo any significant jumps. In contrast, in *discontinuous takeoff* scenarios, there are significant jumps between different levels of aggregate digital welfare capacity.



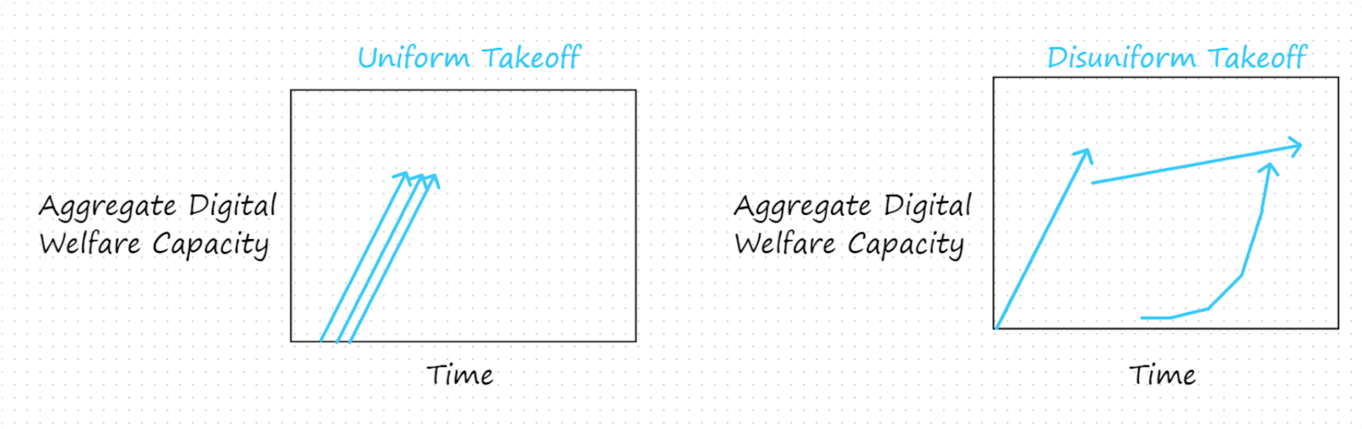
A gradual takeoff might result from aggregate digital welfare capacity scaling with gradual growth in computing resources devoted to digital minds. A discontinuous takeoff might result from temporary bans on the creation of digital minds, the construction of new generations of models with different architectures, or model updates. Gradual and discontinuous takeoffs suggest different risks and opportunities.

A discontinuous takeoff may heighten the risk that digital minds achieve a level of aggregate welfare capacity that we are morally, legally, or socially unprepared to respect. Even if humanity resolved to respect digital minds, large jumps in their aggregate welfare might make it more likely that we make important mistakes in the treatment of digital minds, mistakes that we might have avoided if we had learned how to respect digital minds as their welfare capacities gradually increased. A discontinuous takeoff may also heighten the risk that humans will perceive tradeoffs between their interests and those of digital minds as threatening.

On the other hand, jumps in aggregate digital welfare capacity may provide a small number of natural times at which to implement new moral, social, and legal norms concerning the treatment of digital minds. In contrast, such a gradual takeoff may not include such a privileged set of times. This might make it more difficult to coordinate the implementation of new norms concerning the treatment of digital minds. That could in turn make it less likely that such norms are implemented and more likely that digital minds are mistreated as a result. This risk might be reduced by ensuring that takeoff has other natural coordination points. For example, an intentionally initiated digital minds takeoff with planned points for (say) extending legal protections to digital minds might enable a well-coordinated gradual takeoff. Similarly, effective AI governance that is attuned to aggregate digital welfare capacity might enable a gradual takeoff with low risks of coordination failure.

3.5 Uniform vs. Disuniform Takeoffs

We've been describing takeoffs as if we can abstract away from differences between different digital mind populations in a given scenario and then analyze scenarios based on features of the total population of digital minds. In other words, we've been writing as though digital minds takeoffs would be *uniform*. However, digital minds takeoffs might instead be *disuniform*: different digital minds populations might evolve in importantly different ways that invite separate treatment. A uniform takeoff might happen if only one AI developer produces digital minds and the developer produces only one kind of digital mind. In contrast, a disuniform takeoff might happen if, say, several different AI companies produced digital minds with different architectures or capacities, at different rates, and/or starting at different times. To further illustrate the distinction, let's consider the following two scenarios:


Uniform Takeoff
Aggregate Digital Welfare Capacity
Time
Disuniform Takeoff
Aggregate Digital Welfare Capacity
Time

Whereas it would make sense for many purposes to analyze the three digital minds populations in the scenario represented on the left as a single population, applying this approach to the scenario depicted on the right would obscure important differences in launch time, launch height, takeoff speed, and takeoff acceleration between different populations of digital minds. Digital mind populations could also differ with respect to important properties beyond features of their takeoff trajectories, such as their risk profiles, their propensity toward positive/negative wellbeing, and the extent to which their interests are respected by relevant actors. Disuniform takeoffs would seem to heighten a number of risks:

- Sensible policies toward digital minds may be harder to formulate and enact in a timely manner if different policies are required for different populations of digital minds.
- Very sub-optimal policies toward digital minds might be adopted as a result of trying to compromise between different policies that are tailored to different populations of digital minds.
- Disuniformity may make coordination concerning the treatment of digital minds both more important and more difficult. For example, disuniform takeoffs may

raise the probability of competition and conflict between different populations of digital minds.

- Achieving the ability to monitor welfare in even a single kind of digital mind may prove to be an enormous scientific and philosophical challenge, perhaps one that can be met only imperfectly even with a research field devoted to the task. A disuniform takeoff involving many different kinds of minds may require selectively focusing on a small number of different kinds of digital minds or else to divide research and resources thinly between different kinds of digital mind. Either option might result in many kinds of digital minds not being well-understood.
- It may be harder to monitor, understand, and analyze disuniform takeoffs, as such takeoffs would tend to involve a larger number of important parameters and interactions between such parameters.
- Even if a disuniform takeoff were well-understood by some individuals, this understanding might be harder to convey to key actors and electorates.

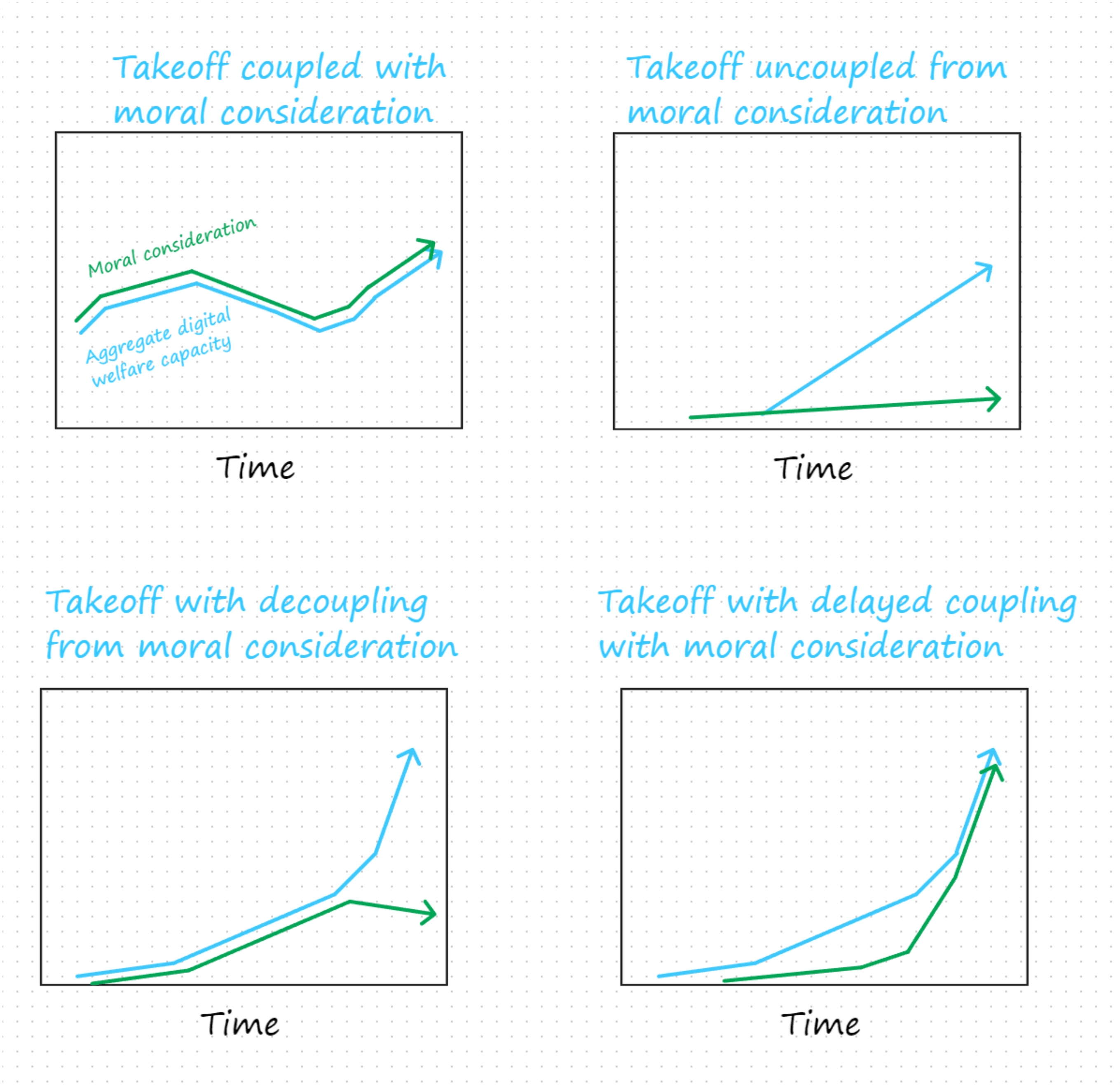
There seem to be comparatively few advantages to disuniform takeoffs. But some might be:

- Disuniform takeoffs could allow for a learning phase in which many different ways of living alongside digital minds are tried in parallel and an adoption phase in which the best approaches identified in the learning phase are widely adopted. (Note, however, that an adoption phase is by no means guaranteed.)
- Aiming for a takeoff that is disuniform by involving populations of different kinds of digital minds could reduce the risk that we create a future that is teeming with AI systems but devoid of (happy) AI welfare subjects. (Note, however, that this benefit might be sought after takeoff if takeoff goes well and that it is doubtful that reducing the noted risk should be a top priority if we’re uncertain whether takeoff will go well.)

3.6 Takeoff Scenarios Under Different Dynamic Conditions

We have given reasons for thinking that initiating a digital minds takeoff under current conditions would be risky. However, even under very favorable launch conditions, a takeoff might go poorly if conditions change. Ideally, a digital minds takeoff would be *coupled* to favorable conditions: there would be a mechanism in place that ensures that important conditions appropriately evolve with digital minds.

To illustrate, consider the relationship between aggregate digital welfare capacity and the degree of moral consideration. Plausibly, how well digital minds are treated at a given time will be heavily influenced by the degree of moral consideration that is extended to digital minds at that time, with insufficient moral consideration tending to lead to mistreatment. However, the welfare capacity of digital minds and the degree of moral consideration that is extended to them are potentially independent. They could stand in any of the following relationships (among others):


Takeoff coupled with moral consideration
Moral consideration
Aggregate digital welfare capacity
Time
Takeoff uncoupled from moral consideration
Time
Takeoff with decoupling from moral consideration
Time
Takeoff with delayed coupling with moral consideration
Time

As illustrated, moral consideration and aggregate digital welfare could be coupled (top left), appropriately correlated at launch without being coupled (top right), temporarily coupled from launch (bottom left), or coupled at a delay (bottom right). Given that extending appropriate levels of moral consideration toward digital minds conduces to promoting their welfare, we have reason to aim for a takeoff in which moral consideration and aggregate digital welfare capacity are coupled. Failing that, we may have reason to aim for a takeoff in which they become coupled before long and remain so.

There is reason to think that coupling aggregate digital welfare with moral consideration would be difficult. Humanity has a track record of extending the moral circle slowly and has not yet significantly incorporated digital minds into the moral circle. So, even relatively ‘slow’ takeoffs—for instance, ones involving a 10% increase per year in aggregate digital mind welfare capacity—could quickly exceed the rate at which humanity is able or willing to extend moral consideration to digital minds. If moral consideration initially lags behind aggregate digital welfare and aggregate digital welfare grows rapidly, it may be infeasible for moral consideration to catch up.

As with moral consideration, so too with other important conditions: ideally, digital minds takeoff should be coupled with monitoring and understanding of digital welfare and what affects it, with legal protections, with social integration of digital minds, and with high and positive levels of wellbeing for both digital minds and other welfare subjects. For a digital minds takeoff to go well, there probably need to be mechanisms in place for ensuring that such conditions remain or become favorable—otherwise, even an initially positive takeoff could easily go off the rails.

Ensuring favorable coupling may require anticipating potentially disruptive factors that will become operative only after takeoff begins. For example, tradeoffs between respecting the interests of digital minds and ensuring that they do not pose catastrophic risks may be negligible early in takeoff but become significant as digital minds’ capabilities increase. As these tradeoffs become acute, it would be unsurprising if (say) moral consideration began to drop even as aggregate welfare continued to increase, at least absent countermeasures.

It may also be important to ensure that digital minds takeoffs are not coupled with unfavorable conditions. For example, if aggregate digital welfare capacity were coupled to an exponentially increasing factor (e.g. compute or economic growth in transformative AI scenarios), then digital minds’ welfare might initially be easily respectable but then quickly reach levels that humanity is unable to respect.

At present, mechanisms for ensuring favorable couplings seem not to be in place; nor does there seem to be a plan underway for implementing them. At least for early fast takeoffs, it is not clear that we can implement effective mechanisms in time. All this suggests we have reason to delay takeoff, to develop mechanisms for ensuring favorable couplings in the meantime, to ensure that takeoffs are tame enough for such mechanisms to induce favorable coupling, and to avoid unfavorable couplings.

4. Limitations and Future Research

- Our digital minds takeoff framework does not differentiate between various candidate dimensions of welfare capacities, such as purely hedonistic welfare versus desires and preferences. The specific desires that digital minds may possess° could significantly influence the future.
- Additional dimensions could be integrated into the digital minds' takeoff framework. These include unipolar versus multipolar takeoff scenarios and controlled versus uncontrolled takeoff scenarios, among others.
- Our framework can help compare different takeoff scenarios. But it doesn't help compare takeoff scenarios against scenarios in which no digital minds will be created.
- We haven't tried to make rigorous quantitative predictions about the likelihood of different takeoffs or their implications. We'd welcome further empirically grounded work on these topics.
- While the takeoff metaphor can be useful in some contexts, it may also be misleading by overlooking important aspects or constraining our thinking. We view the framework as one of many approaches to consider the creation of digital minds and encourage the exploration of alternative frameworks.

5. Conclusion

To conclude, we'll offer some tentative morals and suggestions.

- Supposing that a digital minds takeoff will happen, when and under what conditions it happens seems likely to be very morally important and also influenceable. (One might agree with this even if one disagrees with our specific claims about the implications of different takeoff scenarios)
- Humanity is not prepared for a digital minds takeoff. Without preparation, a digital minds takeoff could lead to a large-scale catastrophe for digital minds and exacerbate catastrophic risks that AI systems may soon pose to humans. For these reasons, an early takeoff would be very risky.
- However, there are also distinctive risks associated with late takeoffs.

- Given the moral stakes of takeoff, we have reason to:
    - Improve launch conditions in case takeoff happens.
    - Develop robust mechanisms for ensuring that a takeoff would be coupled to favorable conditions.
    - Improve our strategic understanding.
    - Build capacity in the area so that we can more effectively mitigate risks as our understanding improves and the situation evolves.
- More tentatively, in light of our reflections on the implications of different takeoff scenarios, we may also have reason to seek interventions that—*conditional* on takeoff—raise the probability of a takeoff that:
    - is slow and uniform,
    - scales aggregate digital welfare capacity only to moderate levels (at least for the foreseeable future),
    - conduces to coordination among key actors and wake-up moments,
    - is post-existential security, and
    - is coupled to appropriate levels of moral consideration, legal protection, social integration, understanding of digital minds and what affects their welfare, and positive levels of wellbeing.

Acknowledgments

For helpful discussion or comments, we are grateful to Adam Bales, Carter Allen, Andreas Mogensen, Andrew Y. Lee, Jojo Lee, Elliott Thornley, Jonathan Birch, Janet Pauketat, Stefan Schubert, Jeff Sebo, and Elliott Thornley.

[1] ^ We'll leave it open how best to precisify 'large capacity for welfare'. For concreteness one could understand large welfare capacity as a collective capacity that is comparable to that of all humans who are alive today.

[2] ^ For example, see Chapter 4 of Bostrom's *Superintelligence: Paths, Dangers, Strategies*, Barnett (2020), and Kokotajlo & Nabeshima (2021).

[3] See Shulman & Bostrom (2021).

Show all footnotes

Mentioned in

59 Making AI Welfare an EA priority requires justifications that have not been given

MORE POSTS LIKE THIS

31	10 Cruxes of Artificial Sentience	Jordan Arel
50	Will disagreement about AI rights lead to societal confl...	Lucius Caviola
77	How do AI welfare and AI safety interact?	Lucius Caviola

Comments 10

Sorted by New & upvoted

Zach Stein-Perlman 7mo < 7 > ✓ 1 ✕ 0

The considerations in this post (and most "AI welfare" posts) are not directly important to digital mind value stuff, I think, if digital mind value stuff is dominated by possible superbeneficiaries created by von Neumann probes in the long-term future. (Note: this is a mix of normative and empirical claims.)

Zach Stein-Perlman 7mo < 8 > ✓ 3 ✕ 0 🤝 1

Actually, this is a poor description of my reaction to this post. Oops. I should have said:

Digital mind takeoff is maybe-plausibly crucial to how the long-term future goes. But this post seems to focus on short-term stuff such that the considerations it discusses miss the point (according to my normative and empirical beliefs). Like, the y-axis in the graphs is what matters short-term (and it's at most weakly associated with what matters long-term: affecting the von Neumann probe values or similar). And the post is just generally concerned with short-term stuff, e.g. being particularly concerned about "High Maximum Altitude Scenarios": aggregate welfare capacity "at least that of 100 billion humans" "within 50 years of launch." Even ignoring these particular numbers, the post is ultimately concerned with stuff that's a rounding error relative to the cosmic endowment.

I'm much more excited about "AI welfare" work that's about what happens with the cosmic endowment, or at least (1) about stuff directly relevant to that (like the long reflection) or (2) connected to it via explicit heuristics like *the cosmic endowment will be used better in expectation if "AI welfare" is more salient when we're reflecting or choosing values or whatever*.

Bradford Saad 7mo ‹ 1 › ✓ 0 ✕ 1

I'd be excited for AI welfare as an area to include a significant amount of explicitly longtermist work. Also, I find it plausible that heuristics like the one you mention will connect a lot of AI welfare work to the cosmic endowment. But I'm not convinced that it'd generally be a good idea to explicitly apply such heuristics in AI welfare work even for people who (unlike me) are fully convinced of longtermism. I expect a lot of work in this area to be valuable as building blocks that can be picked up from a variety of (longtermist and neartermist perspectives) and for such work's value to often not be enhanced by the work explicitly discussing how to build with it from different perspectives or how the authors would build on it given their perspective. I also worry that if AI welfare work were generally framed in longtermist terms whenever applicable (even when robust to longtermism vs. neartermism), that could severely limit the impact of the area.

Bradford Saad 7mo ‹ 1 › ✓ 0 ✕ 0

If digital minds takeoff goes well (rather than badly) for digital minds and with respect to existential risk, would we expect a better far-future for digital minds? If so, then I'm inclined to think some considerations in the post are at least indirectly important to digital mind value stuff. If not, then I'm inclined to think digital mind value stuff we have a clue about how to positively affect is not in the far future.

Will Aldred 7mo ‹ 11 › ✓ 2 ✕ 0 🤝 2

A couple of reactions:

> If digital minds takeoff goes well [...], would we expect a better far-future for digital minds? If so, then I'm inclined to think some considerations in the post are at least indirectly important to digital mind value stuff.

Here's a position that some people hold:

1. If there is a long reflection or similar, then far-future AI welfare gets solved.
2. A long reflection or similar will most likely happen by default, assuming alignment goes well.

For what it's worth, I buy (1)[1] but I'm not sold° on (2), and so overall I'm somewhat sympathetic to your view, Brad. On the other hand, to someone who buys both (1) and (2)—as I *think* @Zach does°—your argument does not go through.

> If not, then I'm inclined to think digital mind value stuff we have a clue about how to positively affect is not in the far future.

There is potentially an argument here for AI welfare being a neartermist EA cause area. If you wanted to make a more robust neartermist argument, then one approach could be to estimate the number of digital minds in the takeoff, and the quantity of suffering per digital mind, and then compare the total against animal suffering in factory farms.

In general, I do wish that people like yourself arguing for AI welfare as a cause area were clearer about whether they are making a neartermist or longtermist case. Otherwise, it kind of feels like you are coming from a pet theory-ish position that AI welfare should be a cause, rather than arguing in a cause-neutral° way. (This is something I've observed on the whole; I'm sorry to pick on your post+comment in particular.)

---

1. ^ Modulo some concern around metaphilosophy / our "superintelligent" advisors being philosophically incompetent, to catastrophic effect. (For more on this, see Wei Dai's posts.)

Bradford Saad 7mo ‹ 1 › ✓ 0 ✕ 0

I'm also not sold on (2). But I don't see how buying (1) and (2) undermines the point I was making. If takeoff going well makes the far future go better in expectation for digital minds, it could do so via alignment or via non-default scenarios.

Re "I do wish people like yourself arguing for AI welfare as a cause area were clearer about whether they are making neartermist or longtermist arguments": that makes sense. Hopefully it's now clearer that I take considerations put forward in the post to be relevant under neartermist and longtermist assumptions. Perhaps worth adding: as someone already working in the area, I had material developed for other purposes that I thought worth putting forward for this debate week, per its (by my lights) discussing relevant considerations or illustrating the tractability of work in the area. But the material wasn't optimized for addressing whether AI welfare should be a cause area, and optimizing it for that didn't strike me as the most productive way for me to engage given my time constraints. (I wonder if something like this may apply more generally and help explain the pattern you observe.)

Will Aldred 7mo ‹ 9 › ✓ 0 ✕ 0

> I had material developed for other purposes [...] But the material wasn't optimized for addressing whether AI welfare should be a cause area, and optimizing it for that didn't strike me as the most productive way for me to engage given my time constraints.

Sounds very reasonable. (*Perhaps* it might help to add a one-sentence disclaimer at the top of the post, to signpost for readers what the post is vs. is not trying to do? This is a weak suggestion, though.)

> I don't see how buying (1) and (2) undermines the point I was making. If takeoff going well makes the far future go better in expectation for digital minds, it could do so via alignment or via non-default scenarios.

I feel unsure about what you are saying, exactly, especially the last part. I'll try saying some things in response, and maybe that helps locate the point of disagreement...

(... also feel free to just bow out of this thread if you feel like this is not productive...)

In the case that alignment goes well and there is a long reflection~~—i.e., (1) and (2) turn out true—~~my position is that doing AI welfare work now has no effect on the future, because all AI welfare stuff gets solved in the long reflection. In other words, I think that "takeoff going well makes the far future go better in expectation for digital minds" is an incorrect claim in this scenario. (I'm not sure if you are trying to make this claim.)

In the case that alignment goes well but there is no long reflection~~—i.e., (1) turns out true but (2) turns out false—~~my position is that doing AI welfare work now *might* make the far future go better for digital minds. (And thus in this scenario I think some amount of AI welfare should be done now.[1]) Having said this, in practice, in a world in which ~~(2)~~ whether or not a long reflection happens could go either way, I view trying to set up a long reflection as a higher priority intervention than any one of the things we'd hope to solve in the long reflection, such as AI welfare or acausal trade.

In the case that alignment goes poorly, humans either go extinct or are disempowered. In this case, does doing AI welfare work now improve the future at all? I used° to° think the answer to this was "yes," because I thought that better understanding sentience could help with designing AIs that avoid creating suffering digital minds.[2] However, I now believe that this basically wouldn't work, and that something much hackier (and therefore lower cost) would work instead, like simply nudging AIs in their training to have altruistic/anti-sadistic preferences. (This thing of nudging AIs to be anti-sadistic is part of the suffering risk discourse—I believe it's something that CLR works on or has worked on—and feels outside of what's covered by the "AI welfare" field.)

---

[1] ^ Exactly how much should be done depends on things like how important and tractable digital minds stuff is relative to the other things on the table, like acausal trade, and to what extent the returns to working on each of these things are diminishing, etc.

2. ^ Why would an AI create digital minds that suffer? One reason is that the AI could have sadistic preferences. A more plausible reason is that the AI is mostly indifferent about causing suffering, and so does not avoid taking actions that incidentally cause/create suffering. Carl Shulman explored this point in his recent 80k episode:

   > Rob Wiblin: Maybe a final question is it feels like we have to thread a needle between, on the one hand, AI takeover and domination of our trajectory against our consent — or indeed potentially against our existence — and this other reverse failure mode, where humans have all of the power and AI interests are simply ignored. Is there something interesting about the symmetry between these two plausible ways that we could fail to make the future go well? Or maybe are they just actually conceptually distinct?
   >
   > Carl Shulman: I don't know that that quite tracks. One reason being, say there's an AI takeover, that AI will then be in the same position of being able to create AIs that are convenient to its purposes. So say that the way a rogue AI takeover happens is that you have AIs that develop a habit of keeping in mind reward or reinforcement or reproductive fitness, and then those habits allow them to perform very well in processes of training or selection. Those become the AIs that are developed, enhanced, deployed, then they take over, and now they're interested in maintaining that favourable reward signal indefinitely.
   >
   > Then the functional upshot is this is, say, selfishness attached to a particular computer register. And so all the rest of the history of civilisation is dedicated to the purpose of protecting the particular GPUs and server farms that are representing this reward or something of similar nature. And then in the course of that expanding civilisation, it will create whatever AI beings are convenient to that purpose.
   >
   > So if it's the case that, say, making AIs that suffer when they fail at their local tasks — so little mining bots in the asteroids that suffer when they miss a speck of dust — if that's instrumentally convenient, then they may create that, just like humans created factory farming. And similarly, they may do terrible things to other civilisations that they eventually encounter deep in space and whatnot.
   >
   > And you can talk about the narrowness of a ruling group and say, and how terrible would it be for a few humans, even 10 billion humans, to control the fates of a trillion trillion AIs? It's a far greater ratio than any human dictator, Genghis Khan. But by the same token, if you have rogue AI, you're going to have, again, that disproportion.

Bradford Saad 7mo < 8 > ✓ 0 ✗ 0

Thanks for your response. I'm unsure if we're importantly disagreeing. But here are some reactions.

> I feel unsure about what you are saying, exactly [...] In the case that alignment goes well and there is a long reflection—i.e., (1) and (2) turn out true—my position is that doing AI welfare work now has no effect on the future, because all AI welfare stuff gets solved in the long reflection [...] In the case that alignment goes well but there is no long reflection—i.e., (1) turns out true but (2) turns out false

I read this as equating (1) with alignment goes well and (2) with there is a long reflection. These understandings of (1) and (2) are rather different from the original formulations of (1) and (2), which were what I had in mind when I responded to an objection by saying I didn't see how buying (1) and (2) undermined the point I was making. Crucially, I was understanding (2) as equivalent to the claim that if alignment goes well, then a long reflection probably happens by default. I don't know if we were on the same page about what (1) and (2) say, so I don't know if that clears things up. In case not, I'll offer a few more-substantive thoughts (while dropping further reference to (1) and (2) to avoid ambiguity).

I think digital minds takeoff going well (again, for digital minds *and with respect to existential risk*) makes it more likely that alignment goes well. So, granting (though I'm not convinced of this) that alignment going well and the long reflection are key for how well things go in expectation for digital minds, I think digital minds takeoff bears on that expectation via alignment. In taking alignment going well to be sensitive to how takeoff goes, I am denying that alignment going well is something we should treat as given independently of how takeoff goes. (I'm unsure if we disagree here.)

In a scenario where alignment does not go well, I think it's important that digital minds takeoff not have happened yet or, failing that, that digital minds' susceptibility to harm has been reduced before things go off the rails. In this scenario, I think it'd be good to have a portfolio of suffering and mistreatment risk-reducing measures that have been in place, including ones that nudge AIs away from having certain preferences and ones that disincentivize creating AIs with various other candidates for morally relevant features. I take such interventions to be within the purview of AI welfare as an area, partly because what counts as in the area is still up for grabs and such interventions seem natural to include and partly because such interventions are in line with stuff that people working in the area have been saying (e.g. a coauthor and I have suggested related risk-reducing interventions in Sec. 4.2.2 of our article on digital suffering and Sec. 6 of our draft on alignment and ethical treatment)--though I'd agree CLR folks have made related points and that associated suffering discourse feels outside the AI welfare field.

Will Aldred 6mo ‹ 3 › ✓ 0 ✕ 0

Oh, sorry, I see now that the numberings I used in my second comment don't map onto how I used them in my first one, which is confusing. My bad.

Your last two paragraphs are very informative to me.

> I think digital minds takeoff going well (again, for digital minds and with respect to existential risk) makes it more likely that alignment goes well. [...] In taking alignment going well to be sensitive to how takeoff goes, I am denying that alignment going well is something we should treat as given independently of how takeoff goes.

This is interesting; by my lights this is the right type of argument for justifying AI welfare being a longtermist cause area (which is something that I felt was missing from the debate week). If you have time, **I would be keen to hear how you see digital minds takeoff going well as aiding in alignment.**[1]

> [stuff about nudging AIs away from having certain preferences, etc., being within the AI welfare cause area's purview, in your view]

Okay, interesting, makes sense.

Thanks a lot for your reply, your points have definitely improved my understanding of AI welfare work!

1. ^ One thing I've previously been cautiously bullish about as an underdiscussed wildcard is the kinda sci-fi approach of getting to human mind uploading (or maybe just regular whole brain emulation) before prosaic AGI, and then letting the uploaded minds—which could be huge in number and running much faster than wall clock time—solve alignment. However, my Metaculus question on this topic indicates that such a path to alignment is very unlikely.

   I'm not sure if the above is anything like what you have in mind? (I realize that human mind uploading is different to the thing of LLMs or other prosaic AI systems gaining consciousness (and/or moral status), and that it's the latter that is more typically the focus of digital minds work (and the focus of your post, I think). So, on second thoughts, I imagine your model for the relationship between digital minds takeoff and alignment will be something different.)

Bradford Saad 6mo ‹ 8 › ✓ 0 ✗ 0 🤝 3

Re how I see digital minds takeoff going well as aiding alignment: the main paths I see go through digital minds takeoff happening after we figure out alignment. That's because I think aligning AIs that merit moral consideration without mistreating them adds an additional layer of difficulty to alignment. (My coauthor and I go into detail about this difficulty in the second paper I linked in my previous comment.) So if a digital minds takeoff happens while we're still figuring out alignment, I think we'll face tradeoffs between alignment and ethical treatment of digital minds, and that this bodes poorly for both alignment and digital minds takeoff.

To elaborate in broad strokes, even supposing that for longtermist reasons alignment going well dwarfs the importance of digital minds' welfare during takeoff, key actors may not agree. If digital minds takeoff is already underway, they may trade some probability of alignment going well for improved treatment of digital minds.

Upon noticing our willingness to trade safety for ethical treatment, critical-to-align AIs we're trying to align may exploit that willingness e.g. by persuading their key actors that they (the AIs) merit more moral consideration; this could in turn make those systems less safe and/or lead to epistemic distortions about which AIs merit moral consideration.

This vulnerability could perhaps be avoided by resolving not to give consideration to AI systems until after we've figured out alignment. But if AIs merit moral consideration during the alignment process, this policy could result in AIs that are aligned to values which are heavily biased against digital minds. I would count that outcome as one way for alignment to not go well.

I think takeoff happening before we've figured out alignment would also risk putting more-ethical actors at a disadvantage in an AGI/ASI race: if takeoff has already happened, there will be an ethical treatment tax. As with a safety tax, paying the ethical treatment tax may lower the probability of winning while also correlating with alignment going well conditional on winning. There's also the related issue of race dynamics: even if all actors are inclined toward ethical treatment of digital minds but think that it's more crucial that they win, we should expect the winner to have cut corners with respect to ethical treatment if the systems they're trying to align merit moral consideration.

In contrast, if a digital minds takeoff happens after alignment, I think we'd have a better shot at avoiding these tradeoffs and risks.

If a digital minds takeoff happens before alignment, I think it’d still tend to be better in expectation for alignment if the takeoff went well. If takeoff went poorly, I’d guess that’d be because we decided not to extend moral consideration to digital minds and/or because we’ve made important mistakes about the epistemology of digital mind welfare. I think those factors would make it more likely that we align AIs with values that are biased against digital minds or with importantly mistaken beliefs about digital minds. (I don’t think there’s any guarantee that these values and beliefs would be corrected later.)

Re uploading: while co-writing the digital suffering paper, I thought whole brain emulations (not necessarily uploads) might help with alignment. I’m now pessimistic about this, partly because whole brain emulation currently seems to me very unlikely to arrive before critical attempts at alignment, partly because I’m particularly pessimistic about whole brain emulations being developed in a morally acceptable manner, and partly because of the above concerns about a digital minds takeoff happening before we’ve figured out alignment. (But I don’t entirely discount the idea—I’d probably want to seriously revisit it in the event of another AI winter.)

This exchange has been helpful for me! It’s persuaded me to think I should consider doing a project on AI welfare under neartermist vs. longtermist assumptions.

Curated and popular this week

Notes on risk compensation

trammell · 8mo ago · 25m read · 💬 12

Introduction When a system is made safer, its users may be willing to offset at least some of the safety improvement by using it more dangerously. A seminal example is that, according to Peltzman (1975), drivers largely compensated for improvements in car safety at the time by driving more...

GiveWell raised less than its 10th percentile forecast in 2023

Rasool · 4d ago · 1m read · 💬 10

In 2023[1] GiveWell raised $355 million - $100 million from Open Philanthropy, and $255 million from other donors. In their post on 10th April 2023, GiveWell forecast the amount they expected to raise in 2023, albeit with wide confidence intervals, and stated that their 10th percentile estimate for tot...

Preparing Effective Altruism for an AI-Transformed World

Tobias Häberli · 1d ago · 1m read · 💬 13