Four visions of Transformative AI success

46

by **Steve Byrnes**

17th Jan 2024

AI Alignment Fieldbuilding | AI Success Models | Research Agendas | AI | Frontpage

Tl;dr

When people work towards making a good future in regards to Transformative AI (TAI), what's the vision of the future that they have in mind and are working towards?

I'll propose four (caricatured) answers that different people seem to give:

- (Vision 1) "Helper AIs",
- (Vision 2) "Autonomous AIs",
- (Vision 3) "Supercharged biological human brains",
- (Vision 4) "Don't build TAI".

For each of these four, I will go through:

- the typical assumptions and ideas that these people seem to typically have in mind;
- potential causes for concern;
- major people, institutions, and research directions associated with this vision.

I'll interject a lot of my own opinions throughout, including a suggestion that, on the current margin, the community should be putting more direct effort into technical work towards contingency-planning for Vision 2.

Warning 1: Oversimplifications. This document is full of oversimplifications and caricatures. But hopefully it's a useful starting point for certain purposes.

Warning 2: Jargon & Unexplained Assumptions. Lots of both; my target audience here is pretty familiar with the AGI safety and alignment literature, and buys into widely-shared assumptions within that literature. But DM me if something seems confusing or dubious, and I'll try to fix it.

Vision 1: “Helper AIs”—AIs doing *specifically* what humans want them to do

1.1 Typical assumptions and ideas

By and large, people in this camp have an assumption that TAI will look, and act, and be trained, much like LLMs, but they’ll work better. They also typically have an assumption of slow takeoff, very high compute requirements for powerful AI, and relatively few big actors who are training and running AIs (but many more actors *using* AI through an API).

There are two common big-picture stories here:

- (Less common story) *Vision 1 is a vision for the long-term future* (example°).
- (More common story) *Vision 1 is a safe way to ultimately get to Vision 2* (*or somewhere else*)—i.e., future people with helper AIs can help solve technical problems related to AI alignment, set up better governance and institutions, or otherwise plan next steps.

1.2 Potential causes for concern

- There’s a risk that somebody makes an *autonomous* (Vision 2 below) ruthlessly-power-seeking AGI, either accidentally or deliberately. We need to either prevent that (presumably through governance), or hope that humans-with-AI-helpers can defend themselves against such AGIs. I’m pretty strongly pessimistic here°, and that is probably my biggest single reason for not buying into this vision. But I’m just one guy, not an expert, and I think reasonable people can disagree.
- Human bad actors will (presumably) be empowered by AI helpers
  - Pessimistic take: It’s really bad if Vladimir Putin (for example) will have a super-smart loyal AI helper.
  - Optimistic take: Well, Vladimir Putin’s opponents will *also* have super-smart loyal AI helpers. So maybe that’s OK!
- “AI slave society” seems kinda bad. Two possible elaborations of that are:
  - “AI slave society *is* in fact bad”; or
  - “Even if AI slave society is *not* in fact bad, at least some humans will *think* that it’s bad. And then those humans will go try to make Vision 2 autonomous AI happen—whether through advocacy and regulation, or by unilateral action.”

- There's no sharp line between the helper AIs of Vision 1 and the truly-autonomous AIs of Vision 2. For example, to what extent do the human supervisors really understand what their AI helpers are doing and how? The less the humans understand, the less we can say that the humans are *really* in control.
    - One issue here is race-to-the-bottom competitive dynamics: if some humans entrust their AIs with more authority to make fast autonomous decisions for complex inscrutable reasons, then those humans will have a competitive advantage over the humans who don't. Thus they will wind up in control of more resources, and in this way, the typical level of human control and supervision may very rapidly drop to zero.
    - Another issue here is that I think people can fool themselves when they try to envision this future. Specifically, it can happen as follows: When you ask yourself a question about AI safety, you say "Oh yes, it will be safe because the AIs will be under extremely close human supervision!" Then an hour later, you ask yourself a question about AI competition and capabilities, and you say "Oh yes, these helper AIs will have all the AI advantages we normally think of, like super-high speed-of-thought, intuitions borne of massive experience, learning, scalability, etc." But really those two answers may be mutually-inconsistent. (Here's a real-life example° where I accused somebody of this kind of misleading equivocation.)

1.3 Who is thinking about this? And if this is your vision, what should you be working on?

- Vision 1 is by-and-large the main vision for people at LLM labs (OpenAI, Anthropic, Conjecture), along with Paul Christiano and OpenPhil, and I think the majority of ML-focused safety / alignment researchers.
- Example technical directions include (I claim) most work on interpretability, scalable oversight, process-based supervision, RLHF, etc.
- Many aspects of contemporary AI governance work is also generally led by people in this camp
    - Examples: Model evaluations, responsible scaling policies, treaties requiring government approval for sufficiently large training runs, incentivizing safety via liability and antitrust law, etc.
- Other work motivated by this kind of vision probably includes Open Agency Architecture°, Comprehensive AI Services°, Bengio's "AI scientists", proof-carrying-code, probably Inverse Reinforcement Learning (Stuart Russell) and most other value

learning work, along with norm learning°, probably "concept extrapolation" (Aligned AI), and much more.

- If we imagine AIs doing what humans *collectively* want, rather than doing what an individual human supervisor wants, then that gets us into some mechanism design challenges. For more discussion see e.g. "democratic inputs to AI" or Critch's discussion of "computational social choice"°.
- Above I mentioned the risk posed by Vision-2-autonomous-ruthlessly-power-seeking AGI in an otherwise-Vision-1 world. Is it a real risk? Can it be managed? How? This is a major crux of disagreement between different thinkers (see intro of my post here°). It would be nice to figure out the answer one way or the other. I haven't seen much work on it. I think there's room for marginal progress here, although we'd probably run into irreducible uncertainty pretty quickly.[1]

Vision 2: "Autonomous AIs"—AIs out in the world, doing whatever they think is best

2.1 Typical assumptions and ideas

By and large, people in this camp have an assumption that TAI will be more in the category of humans, animals, and "RL agents" like AlphaStar. They often talk about AIs that think, figure things out, exhibit plan and foresight, come up with and autonomously implement clever out-of-the-box ways to solve their problems, etc. The AIs are generally assumed to do online learning (a.k.a. "continual learning") as they figure out new things about the world, thus getting more and more competent over time without needing new human-provided training data, just as humans themselves do (individually and in groups). Also, a few people in this camp (not me) think that it's very important in this story that the AI has a robotic body.[2]

As I mentioned in Vision 1 above, there's no sharp line between the helper AIs of Vision 1 and the truly-autonomous AIs of Vision 2. For example, one can imagine a continuum from a 'sycophantic servant AI' that does whatever gets immediate approval from the human; to a 'parent AI' that may ask the human's opinion, and care a lot about it, but also be willing to overrule that opinion in favor of (what it sees as) the human's long-term best interest; to a 'independent AI' that could operate just fine without ever meeting a human in the first place. For clarity, I'll focus discussion on a pretty extreme version of Vision 2.

In that case, an important conceptual distinction (as compared to Vision 1) is related to AI goals:

In Vision 1, there's a pretty straightforward answer of what the AI is supposed to be trying to do—i.e., whatever the human supervisor had in mind, which can be inferred pretty well from some combination of general human data (from which the AI can get context, unspoken assumptions, etc.) and talking to the human in question (from which the AI can get details). The implementation side is by no means straightforward, but in Vision 1, you at least basically know what you're hoping for.

By contrast, in Vision 2, it's head-scratching to even say what the AI is supposed to be doing. We're expecting the AIs to make lots of decisions where "do what the human wants" is not actionable—there might be no human around to ask, and/or not enough time to ask them, and/or the considerations might involve a lot of background knowledge or context that humans don't know, and/or this may be a weird situation where humans would be very unsure (or even mistaken) about what they would want even if those humans *did* understand all the context and consequences. Recall, we're generally expecting the AIs to go invent new science and technology, and build their own idiosyncratic concept-spaces, etc., and *then*, in this new world, which is out-of-distribution relative to all its prior experiences and human data, we generally expect the AIs to continue to make lots of high-context decisions on the fly without necessarily checking in with humans.

So that's a problem. The paths I've heard of for tackling this problem seem to be:[3]

- (A) "Coherent Extrapolated Volition"°;
- (B) "ambitious value learning"°; or
- (C) getting at the deep invariant core of "human values" through neuroscience rather than through human observation and interactions.

The most conceptually-straightforward version of (C) is to start with Whole Brain Emulation (WBE) of unusually decent and upstanding humans, then make it far more competent via speeding it up, tweaking it, adding more virtual cortical neurons, etc. After all, if it's possible for humans to make decisions we're happy about, directly or indirectly, then it's possible in principle for WBEs of those humans to make those same good decisions too; and conversely, if it's *not* possible for humans to make good decisions, directly or indirectly, then we're screwed no matter what.

Another variation on (C) (my favorite!) involves "brain-like AGI"° with (the better parts of) reverse-engineered human social instincts, more on which in 2.3 below.

2.2 Potential causes for concern

- I'm pretty confident that, once there are human-level-ish autonomous AIs doing what they think is best, the entire future of earth-originating life will rapidly (IMO years not decades)[4] stop being under any (biological) human influence (except insofar as the autonomous AIs are motivated to ask the biological humans for their opinions, or to grant them some protected space etc.). Better hope that "what the AIs think is best to do" is also good from a human / moral perspective!
  - This is *directly* bad insofar as it's possible that the AIs will have "bad" values (either initially, or upon reflection, self-modification, creating successors, etc.), and this possibility comprises a single-point-of-failure for everything.
  - This is *procedurally* bad because most existing humans presumably don't want that. It would sure be nice and democratic if those people could have a say!
- Relatedly, humans will stop having any ability to contribute to the economy,[5] and humans themselves will live or die depending on the AIs (more specifically, including both via the AI(s)' individual decisions, and the results of competition / coordination dynamics if this is a multipolar scenario)
  - An optimistic hope is that AIs will feel care and compassion towards humans, so we humans will get a good life, tech advances, and so on. This hope would be loosely in analogy to today's situation in regards to infants, retirees, and pets—i.e., none of those groups can earn money for themselves, or invent things for themselves, but they can do OK thanks to the fact that other people *can* do those things, and feel care and compassion towards those groups.
  - The pessimistic fear is that AI *won't* feel care and compassion towards humans.
  - Another concern goes something like: "We don't want to be outcompeted; we don't want to be the 'pets' or 'helpless infants' of future AI, subject to the whims of their generosity". See also Stuart Russell's discussions of "enfeeblement", or concerns about purposelessness. For my part, purposelessness is pretty low on my list of concerns. For example, retirees today generally feel happy and fulfilled,[6] and likewise, many people find joy and meaning from sorta-pointless activities like climbing mountains, solving crossword puzzles, sports, etc.
- Maybe these AIs will be conscious / sentient. [Note: Some of this bullet point applies to Vision 1 as well.]

  - That's good insofar as the AIs have good lives. Relatedly, if humans *do* wind up extinct, I think it would be *really bad* if we didn't even get the minimal consolation prize of conscious AI successors (Bostrom's "Disneyland with no children").
  - On the other hand, that's bad insofar as the ability to instantiate large amounts of conscious minds on big computers is an s-risk°.
  - This is my controversial opinion, but I strongly expect future powerful AIs to be conscious / sentient°, whether we want that or not. (Relatedly, recall that I'm counting Whole Brain Emulation as an example of Vision 2.)
  - This is also my controversial opinion, but if we're putting some hope on the welfare of future conscious AIs, I think I want them to have a *human-flavored* consciousness—I'd like them to have an innate tendency to care about friendship, compassion, beauty, and so on. This is another reason to hope for either Whole Brain Emulation or brain-like AGI with (some of the) human social instincts.
- As in Vision 1, there's a risk that somebody (perhaps a careless AI?) makes an autonomous ruthlessly-power-seeking AI, and that this AI outcompetes the AIs that care about humans and friendship and so on. Or in a more gradual version of that, there's a risk that progressively-more-ruthless AIs outcompete others. "We" (including the "good" AIs) need to either prevent that somehow, or defend against it.
  - I mentioned in the Vision 1 version that I was very pessimistic° about this genre of concern, but I think in Vision 2 it's not nearly as dire, basically because the "good AIs" are far more powerful than they would be in Vision 1. Specifically, here in Vision 2, the AI(s) can do human-out-of-the-loop autonomous technological development, self-replication, self-improvement, and so on. So hopefully they would be a better match for the "bad AIs", and/or in a better position to forcefully prevent "bad AIs" from getting created in the first place.

2.3 Who is thinking about this? And if this is your vision, what should you be working on?

- Me!! See "brain-like AGI safety"°. My own main research project, described in somewhat more detail here°, is "reverse-engineering human social instincts". I basically posit that human brains involve within-lifetime model-based reinforcement learning, and the reward function for that system involves innate drives related to friendship, compassion, envy, boredom, and many other things that are core to what make humans human, and core to why I'm happier for there to be future generations of humans rather than future generations of arbitrary minds. Anyway, the research project is: Figure out what that reward function is. We probably don't want to directly copy it into AIs in full detail, but it would probably be a good starting point.

  - If we make AI whose "guts" (reward function) overlaps with (the nobler parts of) human innate social drives, then I wouldn't be able to guess what that AI will wind up doing and desiring in any detail, but I'm inclined to feel trust and affinity towards that AI anyway—in a similar way as I feel trust and affinity towards the humans of the next generation, despite likewise not knowing what world they will choose to create, or what they will choose to do with their lives.

- People working on Whole Brain Emulation are also in this category.[7]
- …and that especially includes connectomics! Actually, connectomics is central to *both* of the previous two bullet points (it's essential for Whole Brain Emulation, and it's extremely helpful for reverse-engineering human social instincts). See my Connectomics advocacy post° for much more on this.
- People focused on "ambitious value learning"° AI or Coherent Extrapolated Volition (CEV)°-maximizing AI are generally in this camp. I don't think there are many of them though; most people in value learning / Inverse Reinforcement Learning are more closely aligned with Vision 1, i.e. their "value learning" is not sufficiently "ambitious" to (for example) extrapolate human values into wildly-out-of-distribution societal upheavals, transhumanist transitions, etc. (But there are exceptions—for example, I believe Orthogonal is trying to work towards a CEV-maximizing AI.)
  - That said, there's a decent amount of ongoing "agent foundations research"°, and this is hopefully laying groundwork that could eventually help with ambitious value learning or CEV, among other things.
- Jürgen Schmidhuber and Rich Sutton are among the AI researchers who expect a successor-species AI, but who think that's fine, and thus aren't doing anything in particular to steer that transition, apart from trying to make it happen ASAP. In a similar vein, Robin Hanson frequently talks about *both* AI-as-successor-species (e.g. here), *and* Whole Brain Emulation (*Age Of Em*).
- Building secure simulation sandbox AI testing environments seems like probably a great idea in this vision. For details of why I think that, see Section 4 here° and links therein. (It would also be helpful in Vision 1, but a bit less so I think.)
  - ~~I think Encultured AI is trying to do something related to that?~~ Whoops, nope, they've pivoted.

2.4 Hang on there Steve, this is *your* vision? This is what you actually want?

It's important to distinguish "trying to make this vision happen" from "contingency-planning for this vision". Taking them separately:

- *Should we try to make this vision happen?* I have mixed feelings. On the one hand, I really don't like it—some of the issues mentioned above seem *really bad*, particularly the idea that we're going to make a new intelligent species on the planet despite most humans not wanting that to happen, and also the thing about "single point of failure". On the other hand, maybe the other options are even worse, or not actually viable options in the first place. I guess my opinion is that this vision is probably going to happen, and perhaps without much notice (years not decades), whether it's a good idea or not.
- *Should we plan for the contingency that this kind of thing will happen?* Yes, obviously. Even if you personally really hate this path, we might nevertheless someday find ourselves in the thick of it, so we'd better plan for it and do the best we can.
- *...Yeah but should "we" plan for this contingency? Like, right now? Why not pass the buck to the AI-assisted future humans of Vision 1, as advocated by Paul Christiano, OpenAI°, etc.? Or pass the buck to the enhanced humans of Vision 3, as MIRI has been recently musing°?* My answer: Sure, maybe we can try those buck-passing plans, but we *also* need to be working directly, right now, on contingency-planning for a Vision 2 world. Specifically, we can *hope* to pass the buck to those future Vision 1 or 3 humans, but it may turn out that they'll be only slightly more competent than ourselves, *and* they'll have less time to work on the problem, and indeed they might not appear on the scene in time to help with the problem at all (e.g. see here (Section 4, final bullet point)°).

Vision 3: Supercharged biological human brains (via intelligence-enhancement or merging-with-AI)

3.1 Typical assumptions and ideas

- Two items of fine print:
    - I am defining this vision as centrally involving actual biological neurons. So that means Whole Brain Emulation is Vision 2 (above), *not* Vision 3.

- I'm using the word "intelligence" as shorthand for a broad array of things that contribute to intellectual progress—creativity, insight, work ethic, experience, communication, "scout mindset", and so on.

- Two stories:
  - *Stepping-stone story:* The supercharged human brains will solve the alignment problem or otherwise figure out how to proceed into one of the other three visions.
  - *End-state story:* The supercharged human brains will *become* the superintelligent entities of the future, perhaps by "merging" with AI.

3.2 Potential causes for concern

- The stepping-stone story seems unobjectionable to me as far as it goes, but there's an obvious risk that those "supercharged human brains" will not arrive in time to make any difference for TAI, and/or that they will be only modestly more competent than the traditional human brains of today. So if that's the story, it's really something to be done in parallel with other lines of work that tackle the TAI problem more directly.
- My guess is that the limit of enhanced biological intelligence does not get us anywhere close to competitive with the limit of silicon-chip AIs. Speed is still slow, neuron count is still limited, etc. That's fine in the context of the stepping-stone story—every little bit helps, and we were never expecting to be competitive with future TAI in the first place. But it's a big problem for the end-state story; if you want brains to reign supreme, you need a plan to stop people from making dramatically-more-competent brainless silicon-chip AIs.
- Relatedly, I am very concerned that "merging" is one of those things that sounds great, but only if you don't think about it too hard. I haven't seen any plausible way to flesh it out in detail (or else I haven't understood it).

3.3 Who is thinking about this? And if this is your vision, what should you be working on?

- Example advocacy pieces include Ray Kurzweil books, waitbutwhy on Neuralink, Jed McCaleb's "We have to Upgrade"°, and many more.
  - I think Sam Altman is imagining that Vision 1 Helper AIs will be a stepping stone to Vision 3 "merge" (see his old blog post). (I could be wrong.)

  - MIRI recently expressed enthusiasm° about human intelligence enhancement, but they haven't done anything beyond that, to my knowledge. I think their specific hope is Vision-3-as-a-stepping-stone, and then the more-intelligent future humans will figure out what to do about the TAI problem.
- Work on brain-computer interfaces (BCI) is generally relevant in this vision, including Neuralink (mentioned above), Forest Neurotech, Kernel, and much more.
- There are ideas floating around about making supercharged human brains by embryo selection°, gene therapy in adults°, arguably nootropics, and maybe other stuff; I don't know the details.
- Some aspects of neuroscience, psychology, and connectomics may be relevant here, on the theory that it is probably easier to supercharge a brain, and to interface with it, if you understand how the brain works.

Vision 4: Don't build TAI

4.1 Typical assumptions and ideas

- This camp is an uneasy coalition between *"don't build TAI ever"* and *"don't build TAI yet"*. Both groups are motivated (at least in part) by a concern that TAI could kill everyone (a concern I share). As the saying goes, the idea here is "averting doom by not building the doom machine"°.
- The *"don't build TAI yet"* sub-camp is generally interested in having more time to solve the alignment problem (see here° for more nuance about that).
- The *"don't build TAI ever"* camp is generally just not really into "high-concept sci-fi rigamarole", wherein we transition to a bizarre transhumanist future. Let's stay in the human world and try to make it better, they say.

4.2 Potential causes for concern

- If the idea is to *delay TAI on the margin*, I'm all for it°, other things equal.
  - Other things are definitely not equal: any particular plan or policy would have a whole array of intended and unintended consequences. For example, I have a hot-take opinion° that many popular proposals *purporting* to delay TAI on the margin would in fact unintentionally accelerate it.

  - Anyway, if TAI arrives in 17 years instead of 11, or whatever it is, then I say “hooray, we have more time to prepare”. But we still need to spend that time creating a different plan for TAI success. So this vision would need to be pursued in parallel with “the real plan”, which would be in another category.

- If the idea is to stop TAI *forever*, well I think that’s crazy. How could we know now what AI policies are going to make sense in 50 years—to say nothing of 50,000 years? Also, I for one think friendly superintelligence would be great, cf. “superintelligent AI is necessary for an amazing future°”.
- Moreover, I’m also highly skeptical that “stopping TAI forever” is feasible, even if we wanted it. “Forever” is a very, *very* long time. “Forever” would require much more than just stopping giant training runs. I think it’s probably theoretically possible to run human-level AI on a single consumer GPU°, if only we knew the right algorithms. So (IMO) we would need to eventually either halt all progress on algorithms (which means clamping down hard on things like AI publications, neuroscience publications, PyTorch pull requests, etc.), or send the police from house to house to confiscate consumer GPUs. This strikes me as so extraordinarily unlikely to happen that arguing about it is just a waste of time.

4.3 Who is thinking about this? And if this is your vision, what should you be working on?

- *The populist approach:* You can try to build a popular movement against TAI—see advocacy organizations like Pause AI, stop.ai, and many others.
- *The technocrat approach:* You can reach out to policymakers, draft legislation, track the global flow of chips, etc. Again, I think various organizations are doing that; I don’t know the details.
- *The take-matters-into-my-own-hands approach:* You could build a safe powerful Vision-1-ish AI somehow, and then use it to somehow unilaterally pause global R&D towards TAI. I’m not sure how this is supposed to work in detail, but anyway, this would be the so-called “pivotal act” idea sometimes advocated by MIRI°. I’m not sure anyone is actually working on this, but if they were, the immediate technical details would presumably overlap a lot with Vision 1.

(Thanks Seth Herd, Linda Linsefors, Charlie Steiner, and Adam Marblestone for critical comments on earlier drafts.)

1. ^ One of many challenges is that this kind of scenario planning leans on lots of technical questions about how future AI will work in detail, how competent it will be at different tasks, how much compute it will take to run (both at first, and in the longer term), and so on. It also leans on social questions, like how institutions and individual decision-makers will react in different (unprecedented) circumstances. And it also depends on various aspects of the "tech tree", i.e. what inventions may be invented in the future. These are all really hard questions, so maybe it's no surprise that reasonable people wind up with different opinions.

   By the way, this is a prominent example of my more general rant that there has been insufficient progress and professionalization around thinking through strategies and scenarios of what might happen as we transition into TAI. Part of the problem is that it's really inherently hard and complicated, with a million rabbit-holes and no empirical feedback; and part of the problem is that it sounds like "weird sci-fi stuff", so academics generally won't touch it (besides FHI, to their credit). I'm not really sure how to make this situation better though. (There are a bunch of long TAI-related technical reports from OpenPhil; I have my complaints, but I think that's a good genre.)

2. ^ I strongly expect that future powerful autonomous AIs will be able to *use* teleoperated robot bodies, with very little practice, just as *humans* can use teleoperated robot bodies with very little practice. I don't think it's very important that future AIs *have* robot bodies, in the human or animal sense. For example, lifelong-quadriplegic humans can be remarkably intelligent. More discussion of "embodiment" here°.

3. ^ One can imagine other related scenarios such as "make an AI that wants to set up a Long Reflection and cede power to whatever the result is", or "make an AI that sets up and oversees an atomic communitarian thing". But I think those aren't an *alternative* to (A,B,C) in the text, but rather a *broad strategy* that we might hope the AIs with (A,B,C) type motivations will choose to pursue. After all, you can't just wave a wand and get a Long Reflection; you need to make it happen, in the real world, including setting up appropriate institutions, rules of deliberation, etc., and that would involve the AI making lots of autonomous decisions, long before there is any Long Reflection outputs to defer to. So the AI still needs to have its own motivations that we're happy about.

4. ^ See e.g. Carl Shulman on the possible time-course of AI takeover.

[5] ^ “But what about comparative advantage?” you say. Well, I would point to the example of a moody 7-year-old child in today’s world. *Not only* would nobody hire that kid into their office or high-tech factory, but they would probably pay good money to keep him *out*, because he would only mess stuff up. And if the 7yo could legally found his own company, we would never expect it to get beyond a lemonade stand, given competition from dramatically more capable and experienced adults. So it will be, I claim, with all humans in a world of advanced autonomous AIs, if the humans survive.

[6] ^ I’m not an expert, but see here (including replies) for some references.

[7] ^ In this context, “working on Whole Brain Emulation (WBE)” would include both “making WBE happen” and “arguing about whether WBE is a good idea in the first place”. My own opinion° is that WBE is quite unlikely to happen before AGI (and in particular, very unlikely to happen before having brain-like AGI that is not a WBE of a particular person); but if it did happen, it could be a very useful ingredient in a larger plan, with some care and effort. Others disagree with WBE being desirable in the first place; see e.g. here°.

AI Alignment Fieldbuilding 2 | AI Success Models 2 | Research Agendas 2 | AI 2 | Frontpage

11 comments, sorted by top scoring

[−] **Ryan Greenblatt** 3mo 🔗 < 9 > ✕ 9 ✓

I think your description of vision 1 is likely to give people misleading impressions of what this could plausibly look like or what the people who you cited as pursuing vision 1 are thinking will happen. You disclaim this by noting the doc is oversimplified, but I think various clarifications are quite important in practice.

(It’s possible that you think these misleading impression aren’t that important because from your perspective the main cruxes are in What does it take to defend the world from out-of-control AI°? (But presumably you don’t place total confidence in your views there?))

[*Edit: I think this first paragraph originally came across as more aggressive than I was intending. Sorry. I’ve edited it a bit to tone it down.*]

It seems important to note that the totally amount of autonomy in vision 1 might be extremely large in practice. E.g., AIs might conduct autonomous R&D where some AI instance works on a project for the equivalent of many months without any interaction with a human. (That said I think this system is very likely to be monitored by other AI systems and some actions might be monitored by humans, though it’s plausible that the fraction monitored by humans is very low (e.g. 1%) and long contiguous sequences won’t see any human monitoring.) Levels of autonomy this high might be required for speeding up R&D by large factors (e.g. 30x) due to a combination of serial bottlenecks (meaning that AIs need to serially outspeed humans in many cases) and the obvious argument that a 30x speed up requires AI to automate at least 97% of tasks. (To be clear, I think sometimes when people are imagining vision 1, they aren’t thinking about situations this crazy, but I think they should.)

In fact, I think the level of autonomy between Visions 1 and 2 might be actually similar in practice (because even wild AIs in Vision 2 might want to utilize human labor for some tasks for some transitionary period).

The main difference between vision 1 and visions 2 (assuming vision 1 is working):

- The weights are still on our server and we could turn off the server.
- We can monitor all inputs and outputs from the AI.
- We can continue training the AI and might if we observe undesirable behavior.

> There's no sharp line between the helper AIs of Vision 1 and the truly-autonomous AIs of Vision 2. For example, to what extent do the human supervisors really understand what their AI helpers are doing and how? The less the humans understand, the less we can say that the humans are really in control.

There is also the failure model of deceptive alignment where these AIs are lying in wait for a good opportunity for a treacherous turn. This is a problem even if humans have understood everything they've seen thus far.

> One issue here is race-to-the-bottom competitive dynamics: if some humans entrust their AIs with more authority to make fast autonomous decisions for complex inscrutable reasons, then those humans will have a competitive advantage over the humans who don't. Thus they will wind up in control of more resources, and in this way, the typical level of human control and supervision may very rapidly drop to zero.

Seems like a complicated empirical question. Note that adequately supervising 1% of all queries suffices to rule out a bunch of specific threat models. See auditing failures vs concentrated failures°. Of course, adequate supervision is hard and might be much harder if competitive AIs must perform inscrutable actions which could contain inscrutable danger.

> By and large, people in this camp have an assumption that TAI will look, and act, and be trained, much like LLMs, but they'll work better.

FWIW, I think Paul in particular puts less than 50% on "TAI looks like LLMs" if by that you mean "most of the capabilities come from generative pretraining basically like what we have right now". Short timelines are more likely to look like this though presumably.

[−] **Steve Byrnes** 3mo < 6 > ✕ 1 ✓

That's a very helpful comment, thanks!

Yeah, Vision 1 versus Vision 2 are two caricatures, and as such, they differ along a bunch of axes at once. And I think you're emphasizing on different axes than the ones that seem most salient to me. (Which is fine!)

In particular, maybe I should have focused more on the part where I wrote: "In that case, an important conceptual distinction (as compared to Vision 1) is related to AI goals: *In Vision 1*, there's a pretty straightforward answer of what the AI is supposed to be trying to do… *By contrast, in Vision 2*, it's head-scratching to even say what the AI is supposed to be doing…"

Along *this* axis-of-variation:

- "An AI that can invent a better solar cell, via doing the same sorts of typical human R&D stuff that a human solar cell research team would do" is pretty close to the Vision 1 end of the spectrum, despite the fact that (in a different sense) this AI has massive amounts of "autonomy": all on its own, the AI may rent a lab space, apply for permits, order parts, run experiments using robots, etc.

- The scenario "A bunch of religious fundamentalists build an AI, and the AI notices the error in its programmers' beliefs, and successfully de-converts them" would be much more towards the Vision 2 end of the spectrum—despite the fact that this AI is not very "autonomous" in the going-out-and-doing-things sense. All the AI is doing is thinking, and chatting with its creators. It doesn't have direct physical control of its off-switch, etc.

Why am I emphasizing this axis in particular?

For one thing, I think this axis has practical importance for current research; on the narrow value learning vs ambitious value learning dichotomy°, "narrow" is enough to execute Vision 1, but you need "ambitious" for Vision 2.

For example, if we move from "training by human approval" to "training by human approval after the human has had extensive time to reflect, with weak-AI brainstorming help", then that's a step from Vision 1 towards Vision 2 (i.e. a step from narrow value learning towards ambitious value learning). But my guess is that it's a *pretty small* step towards Vision 2. I don't think it gets us all the way to the AI I mentioned above, the one that will proactively deconvert a religious fundamentalist supervisor who currently has no interest whatsoever in questioning his faith.

For another thing, I think this axis is important for strategy and scenario-planning. For example, if we do Vision 2 really well, it changes the story in regards to "solution to global wisdom and coordination" mentioned in Section 3.2 of my "what does it take" post°.

In other words, I think there are a lot of people (maybe including me) who are wrong about important things, and also not very scout-mindset about those things, such that "AI helpers" wouldn't particularly help, because the person is not asking the AI for its opinion, and would ignore the opinion anyway, or even delete that AI in favor of a more sycophantic one. This is a societal problem, and always has been. One possible view of that problem is: "well, that's fine, we've always muddled through". But if you think there are upcoming VWH-type stuff where we *won't* muddle through (as I tentatively do in regards to ruthlessly-power-seeking AGI), then maybe the only option is a (possibly aggressive) shift in the balance of power towards a scout-mindset-y subpopulation (or at least, a group with more correct beliefs about the relevant topics). That subpopulation could be composed of either humans (cf. "pivotal act"), or of Vision 2 AIs.

Here's another way to say it, maybe. I think you're maybe imagining a dichotomy where either AI is doing what we want it to do (which is normal human stuff like scientific R&D), or the AI is plotting to take over. I'm suggesting that there's a third murky domain where the person wants something that he maybe wouldn't want upon reflection, but where "upon reflection" is kinda indeterminate because he could be manipulated into wanting different things depending on how they're framed. This third domain is important because it contains decisions about politics and society and institutions and ethics and so on. I have concerns that getting an AI to "perform well" in this murky domain is not feasible via a bootstrap thing that starts from the approval of random people; rather, I think a good solution would have to look more like an AI which is *internally* able to do the kinds of reflection and thinking that humans do (but where the AI has the benefit of more knowledge, insight, time, etc.). And that requires that the AI have a certain kind of "autonomy" to reflect on the big picture of what it's doing and why. I think that kind of "autonomy" is different than how you're using the term, but if done well (a big "if"!), it would open up a lot of options.

✓ 2

[−] **Ryan Greenblatt** 3mo < 5 > ✕ 0 ✓

Thanks for the response! I agree that the difference is a difference in emphasis.

[−] **Gunnar Zarncke** 3mo < 7 > ✕ 0 ✓

I think the four scenarios outlined here roughly map to the areas 1, 6, 7, and 8 of the 60+ Possible Futures° post.

[−] **Matthew Barnett** 3mo < 5 > ✕ 7 ✓

> “But what about comparative advantage?” you say. Well, I would point to the example of a not-particularly-bright 7-year-old child in today’s world. *Not only* would nobody hire that kid into their office or factory, but they would probably pay good money to keep him *out*, because he would only mess stuff up.

This is an extremely minor critique given that I'm responding to a footnote, so I hope it doesn't drown out more constructive responses, but I'm actually pretty skeptical that the reason why people don't hire children as workers is because the children would just mess everything up.

I think there are a number of economically valuable physical tasks that most 7-year-old children can perform without messing everything up. For example, one can imagine stocking shelves in stores, small cleaning jobs, and moving lightweight equipment. My thesis here is supported by fact that 7-year-olds were routinely employed to do labor in previous centuries:

> In the 18th century, the arrival of a newborn to a rural family was viewed by the parents as a future beneficial laborer and an insurance policy for old age.[4] At an age as young as 5, a child was expected to help with farm work and other household chores.[5] The agrarian lifestyle common in America required large quantities of hard work, whether it was planting crops, feeding chickens, or mending fences.[6] Large families with less work than children would often send children to another household that could employ them as a maid, servant, or plowboy.[7] Most families simply could not afford the costs of raising a child from birth to adulthood without some compensating labor.

The reason why people don't hire children these days seems more a result of legal and social constraints than the structure of our economy. In modern times, child labor is seen as harmful or even abusive to the child. However, if these legal and social constraints were lifted, arguably most young children in the developed world could be earning wages well above the subsistence level of ~$3/day, making them more productive (in an economic sense) than the majority of workers in pre-modern times.

[−] **Steve Byrnes** 3mo < 4 > ✕ 1 ✓

Thanks. I changed the wording to “moody 7-year-old” and “office or high-tech factory” which puts me on firmer ground I think. :)

I think there have been general increases in productivity across the economy associated with industrialization, automation, complex precise machines, and so on, and those things provide a separate reason (besides legal & social norms as you mentioned) that 7yos are far less employable today than in the 18th century. E.g. I can *easily* imagine a moody 7yo being net useful in a mom & pop artisanal candy shop, but it’s much harder to imagine a moody 7yo being net useful in a modern jelly bean factory.

I think your bringing up “$3/day” gives the wrong idea; I think we should focus on whether the sign is positive or negative. If the sign is positive at all, it’s probably >$3/day. The sign could be negative because they sometimes touch something they’re not supposed to touch, or mess up in other ways, or it could simply be that they bring in extra management overhead greater than their labor contribution. (We’ve all delegated projects where it would have been far less work to just do the project ourselves, right?) E.g. even if the cost to feed and maintain a horse were zero, I would still not expect to see horses being used in a modern construction project.

Anyway, I think I'm on firmer ground when talking about a post-AGI economy, in which case, literally anything that can be done by a human at all, can be automated.

👍 1

[–] **Michele Campolo** 3mo 🔗 < 3 > × 0 ✓

This was a great read, thanks for writing!

Despite the unpopularity of my research° on this forum, I think it's worth saying that I am also working towards Vision 2, with the caveat that autonomy in the real world (e.g. with a robotic body) or on the internet is not necessary: one could aim for an independent-thinker AI° that can do what it thinks is best only by communicating via a chat interface. Depending on what this independent thinker says, different outcomes are possible, including the outcome in which most humans simply don't care about what this independent thinker advocates for, at least initially. This would be an instance of vision 2 with a slow and somewhat human-controlled, instead of rapid, pace of change.

Moreover, I don't know what views they have about autonomy as depicted in Vision 2, but it seems to me that also Shard Theory° and some research bits by Beren Millidge° are to some extent adjacent to the idea of AI which develops its own concept of something being best (and then acts towards it); or, at least, AI which is more human-like in its thinking. Please correct me if I'm wrong.

I hope you'll manage to make progress on brain-like AGI safety! It seems that various research agendas are heading towards the same kind of AI, just from different angles.

[–] **Roko** 3mo 🔗 < 3 > × 0 ✓

Great post. Personally I think the "computational social choice" angle is unerexplored.

I think CSC can gradually morph itself into CEV and that's how we solve AI Goalcraft.

[–] **Steve Byrnes** 3mo 🔗 < 2 > × 0 ✓

> I think CSC can gradually morph itself into CEV and that's how we solve AI Goalcraft.

That sounds lovely if it's true, but I think it's a much more ambitious vision of CSC than people usually have in mind. In particular, CSC (as I understand it) usually takes people's preferences as a given, so if somebody wants something they wouldn't want upon reflection, and maybe they're opposed to doing that reflection because their preferences were always more about signaling etc., well then that's not really in the traditional domain of CSC, but CEV says we ought to sort that out (and I think I agree). More discussion in the last two paragraphs of this comment of mine°.

[–] **Nathan Helm-Burger** 3mo 🔗 < 3 > × 2 ✓

> There's no sharp line between the helper AIs of Vision 1 and the truly-autonomous AIs of Vision 2.

This post seems like it doesn't quite cleave reality at the joints, from how I'm seeing things.

Vision 1 style models can be turned into Vision 2 autonomous models very easily. So, as you say, there′s no sharp line there.

For me, Vision 3 shouldn′t depend on biological neurons. I think it′s more like ′brain-like AGI that is so brain-like that it is basically an accurate whole brain emulation, and thus you can trust it as much as you can trust a human (which isn′t necessarily all that much).″

So again, no sharp line there from my point of view.

Since there are lots of different people in the world with different beliefs and goals, I expect that lots of variations with similarities to #1, #2, and #3 will be active in the world. So anyone who has a hope of just one of the visions coming true needs to include very strict worldwide governance enforcement as part of their vision.

I think my vision is some weird mashup of these. Like, I′m hoping for a powerful set of semi-aligned tool AI (type-1) to assist worldwide enforcement in stamping out dangerous type-2 rogue AI in the hands of bad actors, giving us a temporary safe window in which we can achieve either better alignment of type-1 or type-3 (Bio-enhancement and Whole Brain Emulation).

[−] **Steve Byrnes** 3mo < 6 > ✕ 5 ✓

> Vision 1 style models can be turned into Vision 2 autonomous models very easily

Sure, Vision 1 models can be turned into *dangerous* Vision 2 models, but they can’t be turned into *good* Vision 2 models that we want to have around, unless you solve the different set of problems associated with full-fledged Vision 2. For example, in the narrow value learning vs ambitious value learning dichotomy°, “narrow” is sufficient for Vision 1 to go well, but you need “ambitious” for Vision 2 to go well. Right?

> For me, Vision 3 shouldn′t depend on biological neurons. I think it′s more like ′brain-like AGI that is so brain-like that it is basically an accurate whole brain emulation, and thus you can trust it as much as you can trust a human (which isn′t necessarily all that much).″

I think you’re more focused on “why do I trust the AI (insofar as I trust it)” (e.g. my “two paths” here°), whereas in this post I’m ultimately focused on “what should I be working on (or funding, or whatever) and why”.

Thus, I think “System X does, or does not, involve actual squishy biological neurons” is not only a nice bright line, but it’s also a bright line with great practical importance for what research projects to work on, and what the eventual results will look like, and how the scenarios play out from there. I have lots of reasons for thinking that. E.g. super-ambitious moonshot BCI research is critical for “merging” but only slightly relevant for WBE; conversely measuring human brain connectomes is critical for WBE but only slightly relevant for “merging”. Another example: simbox testing° is useful for WBEs but not “merging”. Also, a WBE would be an extraordinarily powerful system because it can be sped up 100-fold, duplicated, tweaked, and so on, in a way that any system involving actual squishy biological neurons basically can’t (I would argue). And that’s highly relevant to how it fits into longer-term scenarios.

Moderation Log