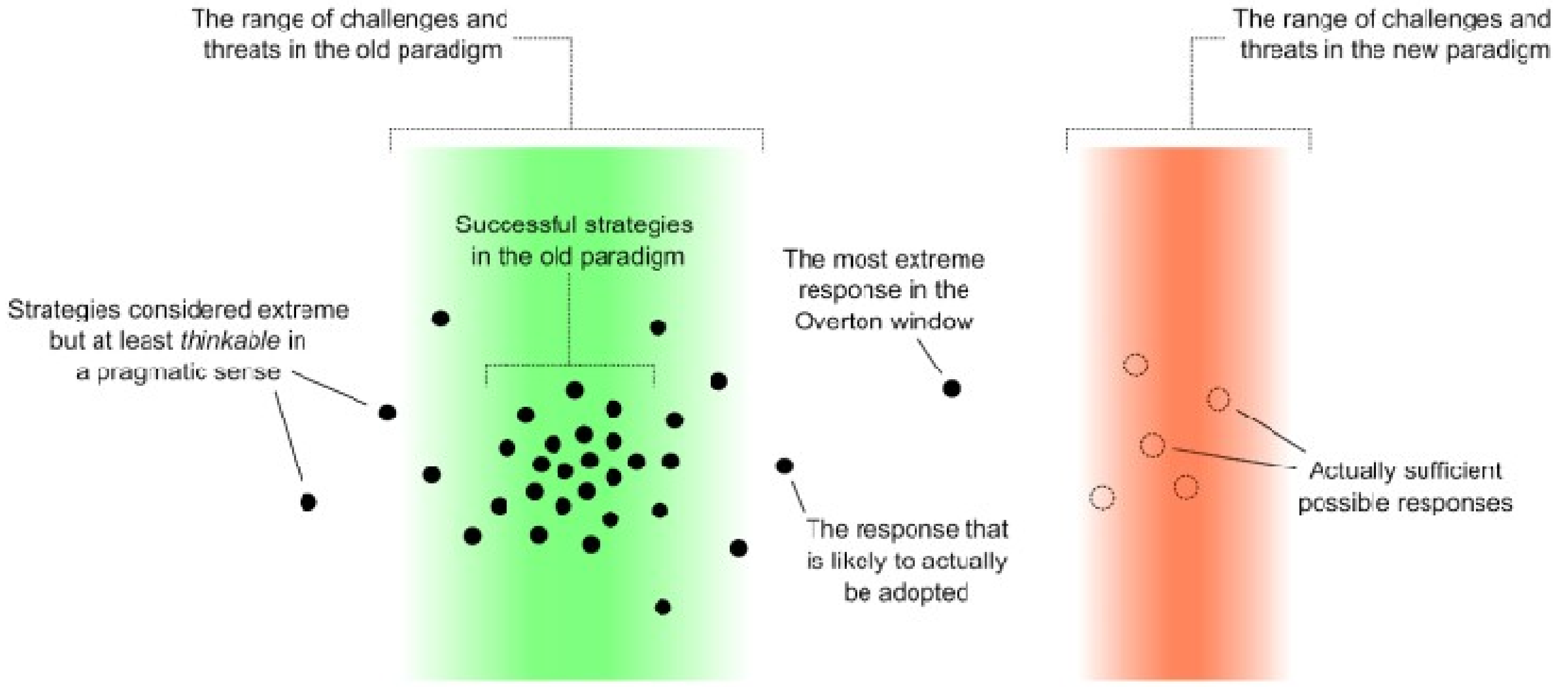

The range of challenges and threats in the old paradigm
The range of challenges and threats in the new paradigm
Successful strategies in the old paradigm
The most extreme response in the Overton window
Strategies considered extreme but at least *thinkable* in a pragmatic sense
Actually sufficient possible responses
The response that is likely to actually be adopted

2022 MIRI Alignment Discussion

Six Dimensions of Operational Adequacy in AGI Projects

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Editor's note: The following is a lightly edited copy of a document written by Eliezer Yudkowsky in November 2017. Since this is a snapshot of Eliezer's thinking at a specific time, we've sprinkled reminders throughout that this is from 2017.

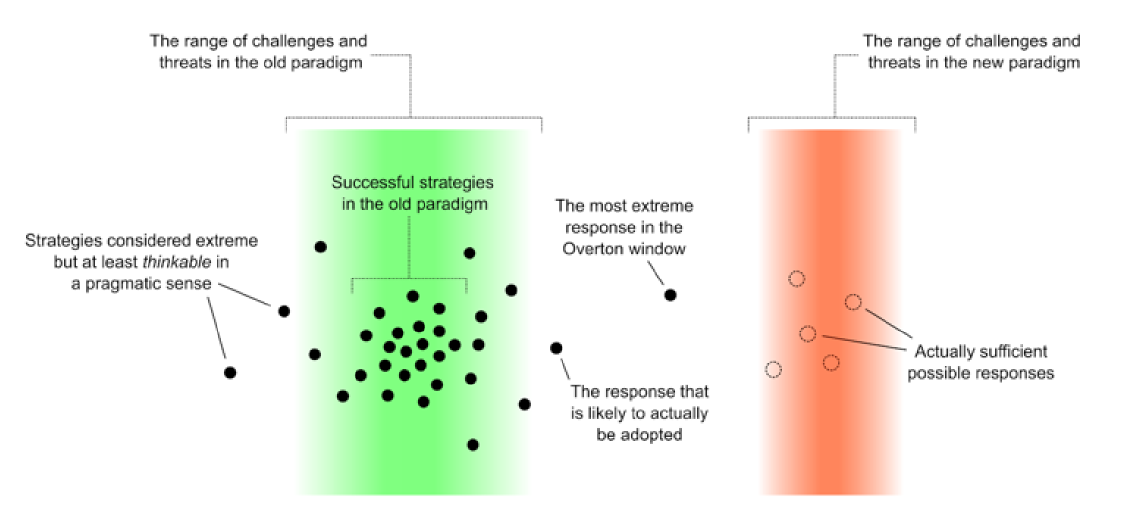
A background note:

It's often the case that people are slow to abandon obsolete playbooks in response to a novel challenge. And AGI is certainly a very novel challenge.

Italian general Luigi Cadorna offers a memorable historical example. In the Isonzo Offensive of World War I, Cadorna lost hundreds of thousands of men in futile frontal assaults against enemy trenches defended by barbed wire and machine guns. As morale plummeted and desertions became epidemic, Cadorna began executing his own soldiers en masse, in an attempt to cure the rest of their "cowardice." The offensive continued for *2.5 years*.

Cadorna made many mistakes, but foremost among them was his refusal to recognize that this war was fundamentally unlike those that had come before. Modern weaponry had forced a paradigm shift, and Cadorna's instincts were not merely miscalibrated—they were systematically broken. No number of small, incremental updates within his obsolete framework would be sufficient to meet the new challenge.

Other examples of this type of mistake include the initial response of the record industry to iTunes and streaming; or, more seriously, the response of most Western governments to COVID-19.


The range of challenges and threats in the old paradigm
The range of challenges and threats in the new paradigm
Successful strategies in the old paradigm
The most extreme response in the Overton window
Strategies considered extreme but at least *thinkable* in a pragmatic sense
The response that is likely to actually be adopted
Actually sufficient possible responses

As usual, the real challenge of reference class forecasting is figuring out which reference class the thing you're trying to model belongs to.

For most problems, rethinking your approach from the ground up is wasteful and unnecessary, because most problems have a similar causal structure to a large number of past cases. When the problem isn’t commensurate with existing strategies, as in the case of AGI, you need a new playbook.

I've sometimes been known to complain, or in a polite way scream in utter terror, that "there is no good guy group in AGI", i.e., if a researcher on this Earth currently wishes to contribute to the common good, there are literally zero projects they can join and no project close to being joinable. In its present version, this document is an informal response to an AI researcher who asked me to list out the qualities of such a "good project".

In summary, a "good project" needs:

- *Trustworthy command:* A trustworthy chain of command with respect to both legal and pragmatic control of the intellectual property (IP) of such a project; a running AGI being included as "IP" in this sense.
- *Research closure:* The organizational ability to *close* and/or *silo* IP to within a trustworthy section and prevent its release by sheer default.
- *Strong opsec:* Operational security adequate to prevent the proliferation of code (or other information sufficient to recreate code within e.g. 1 year) due to e.g. Russian intelligence agencies grabbing the code.
- *Common good commitment:* The project's command and its people must have a credible commitment to both short-term and long-term goodness. Short-term goodness comprises the immediate welfare of present-day Earth; long-term goodness is the achievement of transhumanist astronomical goods.
- *Alignment mindset:* Somebody on the project needs deep enough security mindset plus understanding of AI cognition that they can originate new, deep measures to ensure AGI alignment; and they must be in a position of technical control or otherwise have effectively unlimited political capital. Everybody on the project needs to understand and expect that aligning an AGI will be terrifically difficult and terribly dangerous.
- *Requisite resource levels:* The project must have adequate resources to compete at the frontier of AGI development, including whatever mix of computational resources, intellectual labor, and closed insights are required to produce a 1+ year lead over less cautious competing projects.

I was asked what would constitute "minimal, adequate, and good" performance on each of these dimensions. I tend to divide things sharply into "not adequate" and "adequate" but will try to answer in the spirit of the question nonetheless.

Trustworthy command

Token: Not having pragmatic and legal power in the hands of people who are opposed to the very idea of trying to align AGI, or who want an AGI in every household, or who are otherwise allergic to the *easy* parts of AGI strategy.

E.g.: Larry Page begins with the correct view that cosmopolitan values are good, speciesism is bad, it would be wrong to mistreat sentient beings just because they're implemented in silicon instead of carbon, and so on. But he then proceeds to reject the idea that goals and capabilities are orthogonal, that instrumental strategies are convergent, and that value is complex and fragile. As a consequence, he expects AGI to automatically be friendly, and is liable to object to any effort to align AI as an attempt to keep AI "chained up".

Or, e.g.: As of December 2015, Elon Musk not only wasn’t on board with closure, but apparently wanted to *open-source* superhumanly capable AI.

Elon Musk is not in his own person a majority of OpenAI's Board, but if he can pragmatically sway a majority of that Board then this measure is not being fulfilled even to a token degree.

(Update: Elon Musk stepped down from the OpenAI Board in February 2018.)

Improving: There's a legal contract which says that the Board doesn't control the IP and that the alignment-aware research silo does.

Adequate: The entire command structure including all members of the finally governing Board are fully aware of the difficulty and danger of alignment. The Board will not object if the technical leadership have disk-erasure measures ready in case the Board suddenly decides to try to open-source the AI anyway.

Excellent: Somehow *no* local authority poses a risk of stepping in and undoing any safety measures, etc. I have no idea what incremental steps could be taken in this direction that would not make things worse. If e.g. the government of Iceland suddenly understood how serious things had gotten and granted sanction and security to a project, that would fit this description, but I think that trying to arrange anything like this would probably make things worse globally because of the mindset it promoted.

Closure

Token: It's generally understood organizationally that some people want to keep code, architecture, and some ideas a 'secret' from outsiders, and everyone on the project is okay with this even if they disagree. In principle people aren't being pressed to publish their interesting discoveries if they are obviously capabilities-laden; in practice, somebody always says "but someone else will probably publish a similar idea 6 months later" and acts suspicious of the hubris involved in thinking otherwise, but it remains possible to get away with not publishing at moderate personal cost.

Improving: A subset of people on the project understand why some code, architecture, lessons learned, et cetera must be kept from reaching the general ML community if success is to have a probability significantly greater than zero (because tradeoffs between alignment and capabilities make the challenge unwinnable if there isn’t a project with a reasonable-length lead time). These people have formed a closed silo within the project, with the sanction and acceptance of the project leadership. It's socially okay to be *conservative* about what counts as potentially capabilities-laden thinking, and it's understood that worrying about this is not a boastful act of pride or a trick to get out of needing to write papers.

Adequate: Everyone on the project understands and agrees with closure. Information is siloed whenever not everyone on the project needs to know it.

Reminder: This is a 2017 document.

Opsec

Token: Random people are not allowed to wander through the building.

Improving: Your little brother cannot steal the IP. Stuff is encrypted. Siloed project members sign NDAs.

Adequate: Major governments cannot silently and unnoticeably steal the IP without a nonroutine effort. All project members undergo government-security-clearance-style screening. AGI code is not running on AWS, but in an airgapped server room. There are cleared security guards in the server room.

Excellent: Military-grade or national-security-grade security. (It's hard to see how attempts to get this could avoid being counterproductive, considering the difficulty of obtaining trustworthy command and common good commitment with respect to any entity that can deploy such force, and the effect that trying would have on general mindsets.)

Common good commitment

Token: Project members and the chain of command are not openly talking about how dictatorship is great so long as they get to be the dictator. The project is not directly answerable to Trump or Putin. They say vague handwavy things about how of course one ought to promote democracy and apple pie (applause) and that everyone ought to get some share of the pot o' gold (applause).

Improving: Project members and their chain of command have come out explicitly in favor of being nice to people and eventually building a nice intergalactic civilization. They would release a cancer cure if they had it, their state of deployment permitting, and they don't seem likely to oppose incremental steps toward a postbiological future and the eventual realization of most of the real value at stake.

Adequate: Project members and their chain of command have an explicit commitment to something like coherent extrapolated volition as a long-run goal, AGI tech permitting, and otherwise the careful preservation of values and sentient rights through any pathway of intelligence enhancement. In the short run, they would not do everything that seems to them like a good idea, and would first prioritize not destroying humanity or wounding its spirit with their own hands. (E.g., if Google or Facebook consistently thought like this, they would have become concerned a lot earlier about social media degrading cognition.) Real actual moral humility with policy consequences is a thing.

Alignment mindset

Token: At least some people in command sort of vaguely understand that AIs don't just automatically do whatever the alpha male in charge of the organization wants to have happen. They've hired some people who are at least pretending to work on that in a technical way, not just "ethicists" to talk about trolley problems and which monkeys should get the tasty banana.

Improving: The technical work output by the "safety" group is neither obvious nor wrong. People in command have ordinary paranoia about AIs. They expect alignment to be somewhat difficult and to take some extra effort. They understand that not everything they might like to do, with the first AGI ever built, is equally safe to attempt.

Adequate: The project has realized that building an AGI is *mostly* about aligning it. Someone with full security mindset and deep understanding of AGI cognition as cognition has proven themselves able to originate new deep alignment measures, and is acting as technical lead with effectively unlimited political capital within the organization to make sure the job actually gets done. Everyone expects alignment to be terrifically hard and terribly dangerous and full of invisible bullets whose shadow you have to see before the bullet comes close enough to hit you. They understand that alignment severely constrains architecture and that capability often trades off against transparency. The organization is targeting the minimal AGI doing the least dangerous cognitive work that is required to prevent the next AGI project from destroying the world. The alignment assumptions have been reduced into non-goal-valent statements, have been clearly written down, and are being monitored for their actual truth.

Alignment mindset is *fundamentally* difficult to obtain for a project because Graham's Design Paradox applies. People with only ordinary paranoia may not be able to distinguish the next step up in depth of cognition, and happy innocents cannot distinguish useful paranoia from suits making empty statements about risk and safety. They also tend not to realize what they're missing. This means that there is a horrifically strong default that when you persuade one more research-rich person or organization or government to start a new project, that project *will* have inadequate alignment mindset unless something extra-ordinary happens. I'll be frank and say relative to the present world I think this essentially has to go through trusting me or Nate Soares to actually work, although see below about Paul Christiano. The lack of clear person-independent instructions for how somebody low in this dimension can improve along this dimension is why the difficulty of this dimension is the real killer.

If you insisted on trying this the impossible way, I'd advise that you start by talking to a brilliant computer security researcher rather than a brilliant machine learning researcher.

Resources

Token: The project has a combination of funding, good researchers, and computing power which makes it credible as a beacon to which interested philanthropists can add more funding and other good researchers interested in aligned AGI can join. E.g., OpenAI would qualify as this if it were adequate on the other 5 dimensions.

Improving: The project has size and quality researchers on the level of say Facebook's AI lab, and can credibly compete among the almost-but-not-quite biggest players. When they focus their attention on an unusual goal, they can get it done 1+ years ahead of the general field so long as Demis doesn't decide to do it first. I expect e.g. the NSA would have this level of "resources" if they started playing now but didn't grow any further.

Adequate: The project can get things done with a 2-year lead time on anyone else, and it's not obvious that competitors could catch up even if they focused attention there. DeepMind has a great mass of superior people and unshared tools, and is the obvious candidate for achieving adequacy on this dimension; though they would still need adequacy on other dimensions, and more closure in order to conserve and build up advantages. As I understand it, an adequate resource advantage is explicitly what Demis was trying to achieve, before Elon blew it up, started an openness fad and an arms race, and probably got us all killed. Anyone else trying to be adequate on this dimension would need to pull ahead of DeepMind, merge with DeepMind, or talk Demis into closing more research and putting less effort into unalignable AGI paths.

Excellent: There's a single major project which a substantial section of the research community understands to be The Good Project that good people join, with competition to it deemed unwise and unbeneficial to the public good. This Good Project is at least adequate along all the other dimensions. Its major competitors lack either equivalent funding or equivalent talent and insight. Relative to the present world it would be **extremely difficult** to make any project like this exist with adequately trustworthy command and alignment mindset, and failed attempts to make it exist run the risk of creating still worse competitors developing unaligned AGI.

Unrealistic: There is a single global Manhattan Project which is somehow not answerable to non-common-good command such as Trump or Putin or the United Nations Security Council. It has orders of magnitude more computing power and smart-researcher-labor than anyone else. Something keeps other AGI projects from arising and trying to race with the giant project. The project can freely choose transparency in all transparency-capability tradeoffs and take an extra 10+ years to ensure alignment. The project is at least adequate along all other dimensions.

This is how our distant, surviving cousins are doing it in their Everett branches that diverged centuries earlier towards more competent civilizational equilibria. You **cannot possibly** cause such a project to exist with adequately trustworthy command, alignment mindset, and common-good commitment, and you should therefore not try to make it exist, first because you will simply create a still more dire competitor developing unaligned AGI, and second because if such an AGI could be aligned it would be a hell of an s-risk given the probable command structure. People who are slipping sideways in reality fantasize about being able to do this.

Reminder: This is a 2017 document.

Further Remarks

A project with "adequate" closure and a project with "improving" closure will, if joined, aggregate into a project with "improving" (aka: inadequate) closure where the closed section is a silo within an open organization. Similar remarks apply along other dimensions. The aggregate of a project with NDAs, and a project with deeper employee screening, is a combined project with some unscreened people in the building and hence "improving" opsec.

"Adequacy" on the dimensions of **closure** and **opsec** is based around my mainline-probability scenario where you unavoidably need to spend at least 1 year in a regime where the AGI is not yet alignable on a minimal act that ensures nobody else will destroy the world shortly thereafter, but during that year it's possible to remove a bunch of safeties from the code, shift transparency-capability tradeoffs to favor capability instead, ramp up to full throttle, and immediately destroy the world.

During this time period, leakage of the code to the wider world automatically results in the world being turned into paperclips. Leakage of the code to multiple major actors such as commercial espionage groups or state intelligence agencies seems to me to stand an extremely good chance of destroying the world because at least one such state actor's command will not reprise the alignment debate correctly and each of them will fear the others.

I would also expect that, if key ideas and architectural lessons-learned were to leak from an insufficiently closed project that would otherwise have actually developed alignable AGI, it would be possible to use 10% as much labor to implement a non-alignable world-destroying AGI in a shorter timeframe. The project must be closed *tightly* or everything ends up as paperclips.

"Adequacy" on **common good commitment** is based on my model wherein the first task-directed AGI continues to operate in a regime far below that of a real superintelligence, where many tradeoffs have been made for transparency over capability and this greatly constrains self-modification.

This task-directed AGI is *not* able to defend against true superintelligent attack. It *cannot* monitor other AGI projects in an unobtrusive way that grants those other AGI projects a lot of independent freedom to do task-AGI-ish things so long as they don't create an unrestricted superintelligence. The designers of the first task-directed AGI are *barely* able to operate it in a regime where the AGI doesn't create an unaligned superintelligence inside itself or its environment. Safe operation of the original AGI requires a continuing major effort at supervision. The level of safety monitoring of other AGI projects required would be so great that, if the original operators deemed it good that more things be done with AGI powers, it would be far simpler and safer to do them as additional tasks running on the original task-directed AGI. *Therefore:* Everything to do with invocation of superhuman specialized general intelligence, like superhuman science and engineering, continues to have a single effective veto point.

This is also true in less extreme scenarios where AGI powers can proliferate, but must be very tightly monitored, because no aligned AGI can defend against an unconstrained superintelligence if one is deliberately or accidentally created by taking off too many safeties. Either way, there is a central veto authority that continues to actively monitor and has the power to prevent anyone else from doing anything potentially world-destroying with AGI.

This in turn means that any use of AGI powers along the lines of uploading humans, trying to do human intelligence enhancement, or building a cleaner and more stable AGI to run a CEV, would be subject to the explicit veto of the command structure operating the first task-directed AGI. If this command structure does not favor something like CEV, or vetoes transhumanist outcomes from a transparent CEV, or doesn't allow intelligence enhancement, et cetera, then all future astronomical value can be permanently lost and even s-risks may apply.

A universe in which 99.9% of the sapient beings have no civil rights because way back on Earth somebody decided or *voted* that emulations weren't real people, is a universe plausibly much worse than paperclips. (I would see as self-defeating any argument from democratic legitimacy that ends with almost all sapient beings not being able to vote.)

If DeepMind closed to the silo level, put on adequate opsec, somehow gained alignment mindset within the silo, and allowed trustworthy command of that silo, then in my guesstimation it *might* be possible to save the Earth (we would start to leave the floor of the logistic success curve).

OpenAI seems to me to be further behind than DeepMind along multiple dimensions. OAI is doing significantly better "safety" research, but it is all still inapplicable to serious AGI, AFAIK, even if it's not fake / obvious. I do not think that either OpenAI or DeepMind are out of the basement on the logistic success curve for the alignment-mindset dimension. It's not clear to me from where I sit that the miracle required to grant OpenAI a chance at alignment success is easier than the miracle required to grant DeepMind a chance at alignment success. If Greg Brockman or other decisionmakers at OpenAI are not totally insensible, neither is Demis Hassabis. Both OAI and DeepMind have significant metric distance to cross on Common Good Commitment; this dimension is relatively easier to max out, but it's not maxed out just by having commanders vaguely nodding along or publishing a mission statement about moral humility, nor by a fragile political balance with some morally humble commanders and some morally nonhumble ones. If I had a ton of money and I wanted to get a serious contender for saving the Earth out of OpenAI, I'd probably start by taking however many OpenAI researchers could pass screening and refounding a separate organization out of them, then using that as the foundation for further recruiting.

I have never seen anyone except Paul Christiano try what I would consider to be deep macro alignment work. E.g. if you look at Paul's AGI scheme there is a *global alignment story* with assumptions that can be broken down, and the idea of exact human imitation is a deep one rather than a shallow defense--although I don't think the assumptions have been broken down far enough; but nobody else knows they even ought to be trying to do anything like that. I also think Paul's AGI scheme is orders-of-magnitude too costly and has chicken-and-egg alignment problems. *But* I wouldn't totally rule out a project with Paul in technical command, because I would hold out hope that Paul could follow along with someone else's deep security analysis and understand it in-paradigm even if it wasn't his own paradigm; that Paul would suggest useful improvements and hold the global macro picture to a standard of completeness; and that Paul would take seriously how bad it would be to violate an alignment assumption even if it wasn't an assumption within his native paradigm. Nobody else except myself and Paul is currently in the arena of comparison. If we were both working on the same project it would still have unnervingly few people like that. I think we should try to get more people like this from the pool of brilliant young computer security researchers, not just the pool of machine learning researchers. Maybe that'll fail just as badly, but I want to see it tried.

I doubt that it is possible to produce a written scheme for alignment, or any other kind of fixed advice, that can be handed off to a brilliant programmer with ordinary paranoia and allow them to actually succeed. Some of the deep ideas are going to turn out to be wrong, inapplicable, or just plain missing. Somebody is going to have to notice the unfixable deep problems in advance of an actual blowup, and come up with new deep ideas and not just patches, as the project goes on.

Reminder: This is a 2017 document.

AGI Ruin: A List of Lethalities

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Preamble:

(If you're already familiar with all basics and don't want any preamble, skip ahead to Section B for technical difficulties of alignment proper.)

I have several times failed to write up a well-organized list of reasons why AGI will kill you. People come in with different ideas about why AGI would be survivable, and want to hear different *obviously key* points addressed first. Some fraction of those people are loudly upset with me if the obviously most important points aren't addressed immediately, and I address different points first instead.

Having failed to solve this problem in any good way, I now give up and solve it poorly with a poorly organized list of individual rants. I'm not particularly happy with this list; the alternative was publishing nothing, and publishing this seems marginally more dignified.

Three points about the general subject matter of discussion here, numbered so as not to conflict with the list of lethalities:

-3. I'm assuming you are already familiar with some basics, and already know what 'orthogonality' and 'instrumental convergence' are and why they're true. People occasionally claim to me that I need to stop fighting old wars here, because, those people claim to me, those wars have already been won within the important-according-to-them parts of the current audience. I suppose it's at least true that none of the current major EA funders seem to be visibly in denial about orthogonality or instrumental convergence as such; so, fine. If you don't know what 'orthogonality' or 'instrumental convergence' are, or don't see for yourself why they're true, you need a different introduction than this one.

-2. When I say that alignment is lethally difficult, I am not talking about ideal or perfect goals of 'provable' alignment, nor total alignment of superintelligences on exact human values, nor getting AIs to produce satisfactory arguments about moral dilemmas which sorta-reasonable humans disagree about, nor attaining an absolute certainty of an AI not killing everyone. When I say that alignment is difficult, I mean that in practice, using the techniques we actually have, "please don't disassemble literally everyone with probability roughly 1" is an overly large ask that we are not on course to get. So far as I'm concerned, if you can get a powerful AGI that carries out some pivotal superhuman engineering task, with a less than fifty percent change of killing more than one billion people, I'll take it. Even smaller chances of killing even fewer people would be a nice luxury, but if you can get as incredibly far as "less than roughly certain to kill everybody", then you can probably get down to under a 5% chance with only slightly more effort. Practically all of the difficulty is in getting to "less than certainty of killing literally everyone". Trolley problems are not an interesting subproblem in all of this; if there are any survivors, you solved alignment. At this point, I no longer care how it works, I don't care how you got there, I am cause-agnostic about whatever methodology you used, all I am looking at is prospective results, all I want is that we have justifiable cause to believe of a pivotally useful AGI 'this will not kill literally everyone'. Anybody telling you I'm asking for stricter 'alignment' than this has failed at reading comprehension. The big ask from AGI alignment, the basic challenge I am saying is too difficult, is to obtain by any strategy whatsoever a significant chance of there being any survivors.

-1. None of this is about anything being impossible in principle. The metaphor I usually use is that if a textbook from one hundred years in the future fell into our hands, containing all of the simple ideas *that actually work robustly in practice,* we could probably build an aligned superintelligence in six months. For people schooled in machine learning, I use as my metaphor the difference between ReLU activations and sigmoid activations. Sigmoid activations are complicated and fragile, and do a terrible job of transmitting gradients through many layers; ReLUs are incredibly simple (for the unfamiliar, the activation function is literally max(x, 0)) and work much better. Most neural networks for the first decades of the field used sigmoids; the idea of ReLUs wasn't discovered, validated, and popularized until decades later. What's lethal is that we do not *have* the Textbook From The Future telling us all the simple solutions that actually in real life just work and are robust; we're going to be doing everything with metaphorical sigmoids on the first critical try. No difficulty discussed here about AGI alignment is claimed by me to be impossible - to merely human science and engineering, let alone in principle - if we had 100 years to solve it using unlimited retries, the way that science *usually* has an unbounded time budget and unlimited retries. This list of lethalities is about things *we are not on course to solve in practice in time on the first critical try;* none of it is meant to make a much stronger claim about things that are *impossible in principle.*

That said:

Here, from my perspective, are some different true things that could be said, to contradict various false things that various different people seem to believe, about why AGI would be survivable on anything remotely remotely resembling the current pathway, or any other pathway we can easily jump to.

Section A:

This is a very lethal problem, it has to be solved one way or another, it has to be solved at a minimum strength and difficulty level instead of various easier modes that some dream about, we do not have any visible option of 'everyone' retreating to only solve safe weak problems instead, and failing on the first really dangerous try is fatal.

1. Alpha Zero blew past all accumulated human knowledge about Go after a day or so of self-play, with no reliance on human playbooks or sample games. Anyone relying on "well, it'll get up to human capability at Go, but then have a hard time getting past that because it won't be able to learn from humans any more" would have relied on vacuum. **AGI will not be upper-bounded by human ability or human learning speed**. **Things much smarter than human would be able to learn from less evidence than humans require** to have ideas driven into their brains; there are theoretical upper bounds here, but those upper bounds seem very high. (Eg, each bit of information that couldn't already be fully predicted can eliminate at most half the probability mass of all hypotheses under consideration.) It is not naturally (by default, barring intervention) the case that everything takes place on a timescale that makes it easy for us to react.

2. **A cognitive system with sufficiently high cognitive powers, given any medium-bandwidth channel of causal influence, will not find it difficult to bootstrap to overpowering capabilities independent of human infrastructure.** The concrete example I usually use here is nanotech, because there's been pretty detailed analysis of what definitely look like physically attainable lower bounds on what should be possible with nanotech, and those lower bounds are sufficient to carry the point. My lower-bound model of "how a sufficiently powerful intelligence would kill everyone, if it didn't want to not do that" is that it gets access to the Internet, emails some DNA sequences to any of the many many online firms that will take a DNA sequence in the email and ship you back proteins, and bribes/persuades some human who has no idea they're dealing with an AGI to mix proteins in a beaker, which then form a first-stage nanofactory which can build the actual nanomachinery. (Back when I was first deploying this visualization, the wise-sounding critics said "Ah, but how do you know even a superintelligence could solve the protein folding problem, if it didn't already have planet-sized supercomputers?" but one hears less of this after the advent of AlphaFold 2, for some odd reason.) The nanomachinery builds diamondoid bacteria, that replicate with solar power and atmospheric CHON, maybe aggregate into some miniature rockets or jets so they can ride the jetstream to spread across the Earth's atmosphere, get into human bloodstreams and hide, strike on a timer. **Losing a conflict with a high-powered cognitive system looks at least as deadly as "everybody on the face of the Earth suddenly falls over dead within the same second".** (I am using awkward constructions like 'high cognitive power' because standard English terms like 'smart' or 'intelligent' appear to me to function largely as status synonyms. 'Superintelligence' sounds to most people like 'something above the top of the status hierarchy that went to double college', and they don't understand why that would be all that dangerous? Earthlings have no word and indeed no standard native concept that means 'actually useful cognitive power'. A large amount of failure to panic sufficiently, seems to me to stem from a lack of appreciation for the incredible potential lethality of this thing that Earthlings as a culture have not named.)

3. **We need to get alignment right on the 'first critical try'** at operating at a 'dangerous' level of intelligence, where **unaligned operation at a dangerous level of intelligence kills everybody on Earth and then we don't get to try again**. This includes, for example: (a) something smart enough to build a nanosystem which has been explicitly authorized to build a nanosystem; or (b) something smart enough to build a nanosystem and also smart enough to gain unauthorized access to the Internet and pay a human to put together the ingredients for a nanosystem; or (c) something smart enough to get unauthorized access to the Internet and build something smarter than itself on the number of machines it can hack; or (d) something smart enough to treat humans as manipulable machinery and which has any authorized or unauthorized two-way causal channel with humans; or (e) something smart enough to improve itself enough to do (b) or (d); etcetera. We can gather all sorts of information beforehand *from less powerful systems that will not kill us if we screw up operating them;* but once we are running more powerful systems, we can no longer update on sufficiently catastrophic errors. This is where practically all of the real lethality comes from, that we have to get things right on the first sufficiently-critical try. If we had unlimited retries - if every time an AGI destroyed all the galaxies we got to go back in time four years and try again - we would in a hundred years figure out which bright ideas actually worked. Human beings can figure out pretty difficult things over time, when they get lots of tries; when a failed guess kills literally everyone, that is harder. That we have to get a bunch of key stuff right *on the first try* is where most of the lethality really and ultimately comes from; likewise the fact that no authority is here to tell us a list of what exactly is 'key' and will kill us if we get it wrong. (One remarks that most people are so absolutely and flatly unprepared by their 'scientific' educations to challenge pre-paradigmatic puzzles with no scholarly authoritative supervision, that they do not even realize how much harder that is, or how incredibly lethal it is to demand getting that right on the first critical try.)

4. **We can't just "decide not to build AGI"** because GPUs are everywhere, and knowledge of algorithms is constantly being improved and published; 2 years after the leading actor has the capability to destroy the world, 5 other actors will have the capability to destroy the world. **The given lethal challenge is to solve within a time limit,** driven by the dynamic in which, over time, increasingly weak actors with a smaller and smaller fraction of total computing power, become able to build AGI and destroy the world. Powerful actors all refraining in unison from doing the suicidal thing just delays this time limit - it does not lift it, unless computer hardware and computer software progress are both brought to complete severe halts across the whole Earth. The current state of this cooperation to have every big actor refrain from doing the stupid thing, is that at present some large actors with a lot of researchers and computing power are led by people who vocally disdain all talk of AGI safety (eg Facebook AI Research). Note that needing to solve AGI alignment *only* within a time limit, but with unlimited safe retries for rapid experimentation on the full-powered system; or *only* on the first critical try, but with an unlimited time bound; would both be terrifically humanity-threatening challenges by historical standards *individually*.

5. **We can't just build a very weak system**, which is less dangerous because it is so weak, and declare victory; because later there will be more actors that have the capability to build a stronger system and one of them will do so. I've also in the past called this the 'safe-but-useless' tradeoff, or 'safe-vs-useful'. People keep on going "why don't we only use AIs to do X, that seems safe" and the answer is almost always either "doing X in fact takes very powerful cognition that is not passively safe" or, even more commonly, "because restricting yourself to doing X will not prevent Facebook AI Research from destroying the world six months later". If all you need is an object that doesn't do dangerous things, you could try a sponge; a sponge is very passively safe. Building a sponge, however, does not prevent Facebook AI Research from destroying the world six months later when they catch up to the leading actor.

6. **We need to align the performance of some large task, a 'pivotal act' that prevents other people from building an unaligned AGI that destroys the world.** While the number of actors with AGI is few or one, they must execute some "pivotal act", strong enough to flip the gameboard, using an AGI powerful enough to do that. It's not enough to be able to align a *weak* system - we need to align a system that can do some single *very large thing.* The example I usually give is "burn all GPUs". This is not what I think you'd actually want to do with a powerful AGI - the nanomachines would need to operate in an incredibly complicated open environment to hunt down all the GPUs, and that would be needlessly difficult to align. However, all known pivotal acts are currently outside the Overton Window, and I expect them to stay there. So I picked an example where if anybody says "how dare you propose burning all GPUs?" I can say "Oh, well, I don't *actually* advocate doing that; it's just a mild overestimate for the rough power level of what you'd have to do, and the rough level of machine cognition required to do that, in order to prevent somebody else from destroying the world in six months or three years." (If it wasn't a mild overestimate, then 'burn all GPUs' would actually be the minimal pivotal task and hence correct answer, and I wouldn't be able to give that denial.) Many clever-sounding proposals for alignment fall apart as soon as you ask "How could you use this to align a system that you could use to shut down all the GPUs in the world?" because it's then clear that the system can't do something that powerful, or, if it can do that, the system wouldn't be easy to align. A GPU-burner is also a system powerful enough to, and purportedly authorized to, build nanotechnology, so it requires operating in a dangerous domain at a dangerous level of intelligence and capability; and this goes along with any non-fantasy attempt to name a way an AGI could change the world such that a half-dozen other would-be AGI-builders won't destroy the world 6 months later.

7. The reason why nobody in this community has successfully named a 'pivotal weak act' where you do something weak enough with an AGI to be passively safe, but powerful enough to prevent any other AGI from destroying the world a year later - and yet also we can't just go do that right now and need to wait on AI - is that *nothing like that exists*. There's no reason why it should exist. There is not some elaborate clever reason why it exists but nobody can see it. It takes a lot of power to do something to the current world that prevents any other AGI from coming into existence; nothing which can do that is passively safe in virtue of its weakness. If you can't solve the problem right now (which you can't, because you're opposed to other actors who don't want to be solved and those actors are on roughly the same level as you) then you are resorting to some cognitive system that can do things you could not figure out how to do yourself, that you were not *close* to figuring out because you are not *close* to being able to, for example, burn all GPUs. Burning all GPUs would *actually* stop Facebook AI Research from destroying the world six months later; weaksauce Overton-abiding stuff about 'improving public epistemology by setting GPT-4 loose on Twitter to provide scientifically literate arguments about everything' will be cool but will not actually prevent Facebook AI Research from destroying the world six months later, or some eager open-source collaborative from destroying the world a year later if you manage to stop FAIR specifically. **There are no pivotal weak acts**.

8. **The best and easiest-found-by-optimization algorithms for solving problems we want an AI to solve, readily generalize to problems we'd rather the AI not solve**; you can't build a system that only has the capability to drive red cars and not blue cars, because all red-car-driving algorithms generalize to the capability to drive blue cars.

9. The builders of a safe system, by hypothesis on such a thing being possible, would need to operate their system in a regime where it has the *capability* to kill everybody or make itself even more dangerous, but has been successfully designed to not do that. **Running AGIs doing something pivotal are not passively safe,** they're the equivalent of nuclear cores that require actively maintained design properties to not go supercritical and melt down.

Section B:

Okay, but as we all know, modern machine learning is like a genie where you just give it a wish, right? Expressed as some mysterious thing called a 'loss function', but which is basically just equivalent to an English wish phrasing, right? And then if you pour in enough computing power you get your wish, right? So why not train a giant stack of transformer layers on a dataset of agents doing nice things and not bad things, throw in the word 'corrigibility' somewhere, crank up that computing power, and get out an aligned AGI?

Section B.1: The distributional leap.

10. You can't train alignment by running lethally dangerous cognitions, observing whether the outputs kill or deceive or corrupt the operators, assigning a loss, and doing supervised learning. **On anything like the standard ML paradigm, you would need to somehow generalize optimization-for-alignment you did in safe conditions, across a big distributional shift to dangerous conditions**. (Some generalization of this seems like it would have to be true even outside that paradigm; you wouldn't be working on a live unaligned superintelligence to align it.) This alone is a point that is sufficient to kill a lot of naive proposals from people who never did or could concretely sketch out any specific scenario of what training they'd do, in order to align what output - which is why, of course, they never concretely sketch anything like that. **Powerful AGIs doing dangerous things that will kill you if misaligned, must have an alignment property that generalized far out-of-distribution from safer building/training operations that didn't kill you.** This is where a huge amount of lethality comes from on anything remotely resembling the present paradigm. Unaligned operation at a dangerous level of intelligence*capability will kill you; so, if you're starting with an unaligned system and labeling outputs in order to get it to learn alignment, the training regime or building regime must be operating at some lower level of intelligence*capability that is passively safe, where its currently-unaligned operation does not pose any threat. (Note that anything substantially smarter than you poses a threat given *any* realistic level of capability. Eg, "being able to produce outputs that humans look at" is probably sufficient for a generally much-smarter-than-human AGI to navigate its way out of the causal systems that are humans, especially in the real world where somebody trained the system on terabytes of Internet text, rather than somehow keeping it ignorant of the latent causes of its source code and training environments.)

11. If cognitive machinery doesn't generalize far out of the distribution where you did tons of training, it can't solve problems on the order of 'build nanotechnology' where it would be too expensive to run a million training runs of failing to build nanotechnology. There is no pivotal act this weak; **there's no known case where you can entrain a safe level of ability on a safe environment where you can cheaply do millions of runs, and deploy that capability to save the world** and prevent the next AGI project up from destroying the world two years later. Pivotal weak acts like this aren't known, and not for want of people looking for them. So, again, you end up needing alignment to generalize way out of the training distribution - not just because the training environment needs to be safe, but because the training environment probably also needs to be *cheaper* than evaluating some real-world domain in which the AGI needs to do some huge act. You don't get 1000 failed tries at burning all GPUs - because people will notice, even leaving out the consequences of capabilities success and alignment failure.

12. **Operating at a highly intelligent level is a drastic shift in distribution from operating at a less intelligent level**, opening up new external options, and probably opening up even more new internal choices and modes. Problems that materialize at high intelligence and danger levels may fail to show up at safe lower levels of intelligence, or may recur after being suppressed by a first patch.

13. **Many alignment problems of superintelligence will not naturally appear at pre-dangerous, passively-safe levels of capability**. Consider the internal behavior 'change your outer behavior to deliberately look more aligned and deceive the programmers, operators, and possibly any loss functions optimizing over you'. This problem is one that will appear at the superintelligent level; if, being otherwise ignorant, we guess that it is among the *median* such problems in terms of how *early* it naturally appears in earlier systems, then around *half* of the alignment problems of superintelligence will first naturally materialize *after* that one first starts to appear. Given *correct* foresight of which problems will naturally materialize *later,* one could try to deliberately materialize such problems earlier, and get in some observations of them. This helps to the extent (a) that we actually correctly forecast all of the problems that will appear later, or some superset of those; (b) that we succeed in preemptively materializing a superset of problems that will appear later; and (c) that we can actually solve, in the earlier laboratory that is out-of-distribution for us relative to the real problems, those alignment problems that would be lethal if we mishandle them when they materialize later. Anticipating *all* of the really dangerous ones, and then successfully materializing them, in the correct form for early solutions to generalize over to later solutions, *sounds possibly kinda hard*.

14. **Some problems**, like 'the AGI has an option that (looks to it like) it could successfully kill and replace the programmers to fully optimize over its environment', **seem like their natural order of appearance could be that they first appear only in fully dangerous domains**. Really actually having a *clear* option to brain-level-persuade the operators or escape onto the Internet, build nanotech, and destroy all of humanity - in a way where you're fully clear that you know the relevant facts, and estimate only a not-worth-it low probability of learning something which changes your preferred strategy if you bide your time another month while further growing in capability - is an option that first gets evaluated for real at the point where an AGI fully expects it can defeat its creators. We can try to manifest an echo of that apparent scenario in earlier toy domains. Trying to train by gradient descent against that behavior, in that toy domain, is something I'd expect to produce not-particularly-coherent local patches to thought processes, which would break with near-certainty inside a superintelligence generalizing far outside the training distribution and thinking very different thoughts. Also, programmers and operators themselves, who are used to operating in not-fully-dangerous domains, are operating out-of-distribution when they enter into dangerous ones; our methodologies may at that time break.

15. **Fast capability gains seem likely, and may break lots of previous alignment-required invariants simultaneously.** Given otherwise insufficient foresight by the operators, I'd expect a lot of those problems to appear approximately simultaneously after a sharp capability gain. See, again, the case of human intelligence. We didn't break alignment with the 'inclusive reproductive fitness' outer loss function, immediately after the introduction of farming - something like 40,000 years into a 50,000 year Cro-Magnon takeoff, as was itself running very quickly relative to the outer optimization loop of natural selection. Instead, we got a lot of technology more advanced than was in the ancestral environment, including contraception, in one very fast burst relative to the speed of the outer optimization loop, late in the general intelligence game. We started reflecting on ourselves a lot more, started being programmed a lot more by cultural evolution, and lots and lots of assumptions underlying our alignment in the ancestral training environment broke simultaneously. (People will perhaps rationalize reasons why this abstract description doesn't carry over to gradient descent; eg, “gradient descent has less of an information bottleneck”. My model of this variety of reader has an inside view, which they will label an outside view, that assigns great relevance to some other data points that are *not* observed cases of an outer optimization loop producing an inner general intelligence, and assigns little importance to our one data point actually featuring the phenomenon in question. When an outer optimization loop actually produced general intelligence, it broke alignment after it turned general, and did so relatively late in the game of that general intelligence accumulating capability and knowledge, almost immediately before it turned 'lethally' dangerous relative to the outer optimization loop of natural selection. Consider skepticism, if someone is ignoring this one warning, especially if they are not presenting equally lethal and dangerous things that they say will go wrong instead.)

Section B.2: Central difficulties of outer and inner alignment.

16. Even if you train really hard on an exact loss function, that doesn't thereby create an explicit internal representation of the loss function inside an AI that then continues to pursue that exact loss function in distribution-shifted environments. Humans don't explicitly pursue inclusive genetic fitness; **outer optimization even on a very exact, very simple loss function doesn't produce inner optimization in that direction**. This happens *in practice in real life,* it is what happened in *the only case we know about*, and it seems to me that there are deep theoretical reasons to expect it to happen again: the *first* semi-outer-aligned solutions found, in the search ordering of a real-world bounded optimization process, are not inner-aligned solutions. This is sufficient on its own, even ignoring many other items on this list, to trash entire categories of naive alignment proposals which assume that if you optimize a bunch on a loss function calculated using some simple concept, you get perfect inner alignment on that concept.

17. More generally, a superproblem of 'outer optimization doesn't produce inner alignment' is that **on the current optimization paradigm there is no general idea of how to get particular inner properties into a system, or verify that they're there, rather than just observable outer ones you can run a loss function over.** This is a problem when you're trying to generalize out of the original training distribution, because, eg, the outer behaviors you see could have been produced by an inner-misaligned system that is deliberately producing outer behaviors that will fool you. We don't know how to get any bits of information into the *inner* system rather than the *outer* behaviors, in any systematic or general way, on the current optimization paradigm.

18. **There's no reliable Cartesian-sensory ground truth** (reliable loss-function-calculator) **about whether an output is 'aligned'**, because some outputs destroy (or fool) the human operators and produce a different environmental causal chain behind the externally-registered loss function. That is, if you show an agent a reward signal that's currently being generated by humans, the signal is not *in general* a *reliable perfect ground truth* about *how aligned an action was*, because another way of producing a high reward signal is to deceive, corrupt, or replace the human operators with a different causal system which generates that reward signal. When you show an agent an environmental reward signal, you are not showing it something that is a reliable ground truth about whether the system did the thing you wanted it to do; *even if* it ends up perfectly inner-aligned on that reward signal, or learning some concept that *exactly* corresponds to 'wanting states of the environment which result in a high reward signal being sent', an AGI strongly optimizing on that signal will kill you, because the sensory reward signal was not a ground truth about alignment (as seen by the operators).

19. More generally, **there is no known way to use the paradigm of loss functions, sensory inputs, and/or reward inputs, to optimize anything within a cognitive system to point at particular things within the environment** - to point to *latent events and objects and properties in the environment,* rather than *relatively shallow functions of the sense data and reward.* This isn't to say that nothing in the system’s goal (whatever goal accidentally ends up being inner-optimized over) could ever point to anything in the environment by *accident*. Humans ended up pointing to their environments at least partially, though we've got lots of internally oriented motivational pointers as well. But insofar as the current paradigm works at all, the on-paper design properties say that it only works for aligning on known direct functions of sense data and reward functions. All of these kill you if optimized-over by a sufficiently powerful intelligence, because they imply strategies like 'kill everyone in the world using nanotech to strike before they know they're in a battle, and have control of your reward button forever after'. It just isn't *true* that we know a function on webcam input such that every world with that webcam showing the right things is safe for us creatures outside the webcam. This general problem is a fact about the territory, not the map; it's a fact about the actual environment, not the particular optimizer, that lethal-to-us possibilities exist in some possible environments underlying every given sense input.

20. Human operators are fallible, breakable, and manipulable. **Human raters make systematic errors - regular, compactly describable, predictable errors**. To *faithfully* learn a function from 'human feedback' is to learn (from our external standpoint) an unfaithful description of human preferences, with errors that are not random (from the outside standpoint of what we'd hoped to transfer). If you perfectly learn and perfectly maximize *the referent of* rewards assigned by human operators, that kills them. It's a fact about the territory, not the map - about the environment, not the optimizer - that the *best predictive* explanation for human answers is one that predicts the systematic errors in our responses, and therefore is a psychological concept that correctly predicts the higher scores that would be assigned to human-error-producing cases.

21. There's something like a single answer, or a single bucket of answers, for questions like 'What's the environment really like?' and 'How do I figure out the environment?' and 'Which of my possible outputs interact with reality in a way that causes reality to have certain properties?', where a simple outer optimization loop will straightforwardly shove optimizees into this bucket. When you have a wrong belief, reality hits back at your wrong predictions. When you have a broken belief-updater, reality hits back at your broken predictive mechanism via predictive losses, and a gradient descent update fixes the problem in a simple way that can easily cohere with all the other predictive stuff. In contrast, when it comes to a choice of utility function, there are unbounded degrees of freedom and multiple reflectively coherent fixpoints. Reality doesn't 'hit back' against things that are locally aligned with the loss function on a particular range of test cases, but globally misaligned on a wider range of test cases. This is the very abstract story about why hominids, once they finally started to generalize, generalized their *capabilities* to Moon landings, but their inner optimization no longer adhered very well to the outer-optimization goal of 'relative inclusive reproductive fitness' - even though they were in their ancestral environment optimized very strictly around this one thing and nothing else. This abstract dynamic is something you'd expect to be true about outer optimization loops on the order of both 'natural selection' and 'gradient descent'. The central result: **Capabilities generalize further than alignment once capabilities start to generalize far**.

22. There's a relatively simple core structure that explains why complicated cognitive machines work; which is why such a thing as general intelligence exists and not just a lot of unrelated special-purpose solutions; which is why capabilities generalize after outer optimization infuses them into something that has been optimized enough to become a powerful inner optimizer. The fact that this core structure is simple and relates generically to low-entropy high-structure environments is why humans can walk on the Moon. **There is no analogous truth about there being a simple core of alignment**, especially not one that is *even easier* for gradient descent to find than it would have been for natural selection to just find 'want inclusive reproductive fitness' as a well-generalizing solution within ancestral humans. Therefore, capabilities generalize further out-of-distribution than alignment, once they start to generalize at all.

23. **Corrigibility is anti-natural to consequentialist reasoning**; "you can't bring the coffee if you're dead" for almost every kind of coffee. We (MIRI) tried and failed to find a coherent formula for an agent that would let itself be shut down (without that agent actively trying to get shut down). Furthermore, many anti-corrigible lines of reasoning like this may only first appear at high levels of intelligence.

24. There are two fundamentally different approaches you can potentially take to alignment, which are unsolvable for two different sets of reasons; therefore, **by becoming confused and ambiguating between the two approaches, you can confuse yourself about whether alignment is necessarily difficult**. The first approach is to build a CEV-style Sovereign which wants exactly what we extrapolated-want and is therefore safe to let optimize all the future galaxies without it accepting any human input trying to stop it. The second course is to build corrigible AGI which doesn't want exactly what we want, and yet somehow fails to kill us and take over the galaxies despite that being a convergent incentive there.

1. The first thing generally, or CEV specifically, is unworkable because **the complexity of what needs to be aligned or meta-aligned for our Real Actual Values is far out of reach for our FIRST TRY at AGI** . Yes I mean specifically that the *dataset, meta-learning algorithm, and what needs to be learned,* is far out of reach for our first try. It's not just non-hand-codable, it is *unteachable* on-the-first-try because *the thing you are trying to teach is too weird and complicated.*
2. The second thing looks unworkable (less so than CEV, but still lethally unworkable) because **corrigibility runs *actively counter* to instrumentally convergent behaviors** within a core of general intelligence (the capability that generalizes far out of its original distribution). You're not trying to make it have an opinion on something the core was previously neutral on. You're trying to take a system implicitly trained on lots of arithmetic problems until its machinery started to reflect the common coherent core of arithmetic, and get it to say that as a special case 222 + 222 = 555. You can maybe train something to do this in a particular training distribution, but it's incredibly likely to break when you present it with new math problems far outside that training distribution, on a system which successfully generalizes capabilities that far at all.

Section B.3: Central difficulties of *sufficiently good and useful* transparency / interpretability.

25. **We've got no idea what's actually going on inside the giant inscrutable matrices and tensors of floating-point numbers**. Drawing interesting graphs of where a transformer layer is focusing attention doesn't help if the question that needs answering is "So was it planning how to kill us or not?"

26. Even if we did know what was going on inside the giant inscrutable matrices while the AGI was still too weak to kill us, this would just result in us dying with more dignity, if DeepMind refused to run that system and let Facebook AI Research destroy the world two years later. **Knowing that a medium-strength system of inscrutable matrices is planning to kill us, does not thereby let us build a high-strength system of inscrutable matrices that isn't planning to kill us**.

27. When you explicitly optimize against a detector of unaligned thoughts, you're partially optimizing for more aligned thoughts, and partially optimizing for unaligned thoughts that are harder to detect. **Optimizing against an interpreted thought optimizes against interpretability**.

28. The AGI is smarter than us in whatever domain we're trying to operate it inside, so we cannot mentally check all the possibilities it examines, and we cannot see all the consequences of its outputs using our own mental talent. **A powerful AI searches parts of the option space we don't, and we can't foresee all its options**.

29. The outputs of an AGI go through a huge, not-fully-known-to-us domain (the real world) before they have their real consequences. **Human beings cannot inspect an AGI's output to determine whether the consequences will be good**.

30. Any pivotal act that is not something we can go do right now, will take advantage of the AGI figuring out things about the world we don't know so that it can make plans we wouldn't be able to make ourselves. It knows, at the least, the fact we didn't previously know, that some action sequence results in the world we want. Then humans will not be competent to use their own knowledge of the world to figure out all the results of that action sequence. An AI whose action sequence you can fully understand all the effects of, before it executes, is much weaker than humans in that domain; you couldn't make the same guarantee about an unaligned human as smart as yourself and trying to fool you. **There is no pivotal output of an AGI that is humanly checkable and can be used to safely save the world but only after checking it**; this is another form of pivotal weak act which does not exist.

31. A strategically aware intelligence can choose its visible outputs to have the consequence of deceiving you, including about such matters as whether the intelligence has acquired strategic awareness; **you can't rely on behavioral inspection to determine facts about an AI which that AI might want to deceive you about**. (Including how smart it is, or whether it's acquired strategic awareness.)

32. Human thought partially exposes only a partially scrutable outer surface layer. Words only trace our real thoughts. Words are not an AGI-complete data representation in its native style. The underparts of human thought are not exposed for direct imitation learning and can't be put in any dataset. **This makes it hard and probably impossible to train a powerful system entirely on imitation of human words or other human-legible contents**, which are only impoverished subsystems of human thoughts; ***unless* that system is powerful enough to contain inner intelligences figuring out the humans**, and at that point it is no longer really working as imitative human thought.

33. **The AI does not think like you do**, the AI doesn't have thoughts built up from the same concepts you use, it is utterly alien on a staggering scale. Nobody knows what the hell GPT-3 is thinking, not *only* because the matrices are opaque, but because the *stuff within that opaque container* is, very likely, incredibly alien - nothing that would translate well into comprehensible human thinking, even if we could see past the giant wall of floating-point numbers to what lay behind.

Section B.4: Miscellaneous unworkable schemes.

34. **Coordination schemes between superintelligences are not things that humans can participate in** (eg because humans can't reason reliably about the code of superintelligences); a "multipolar" system of 20 superintelligences with different utility functions, plus humanity, has a natural and obvious equilibrium which looks like "the 20 superintelligences cooperate with each other but not with humanity".

35. Schemes for playing "different" AIs off against each other stop working if those AIs advance to the point of being able to coordinate via reasoning about (probability distributions over) each others' code. **Any system of sufficiently intelligent agents can probably behave as a single agent, even if you imagine you're playing them against each other.** Eg, if you set an AGI that is secretly a paperclip maximizer, to check the output of a nanosystems designer that is secretly a staples maximizer, then even if the nanosystems designer is not able to deduce what the paperclip maximizer really wants (namely paperclips), it could still logically commit to share half the universe with any agent checking its designs if those designs were allowed through, *if* the checker-agent can verify the suggester-system's logical commitment and hence logically depend on it (which excludes human-level intelligences). Or, if you prefer simplified catastrophes without any logical decision theory, the suggester could bury in its nanosystem design the code for a new superintelligence that will visibly (to a superhuman checker) divide the universe between the nanosystem designer and the design-checker.

36. What makes an air conditioner 'magic' from the perspective of say the thirteenth century, is that even if you correctly show them the design of the air conditioner in advance, they won't be able to understand from seeing that design why the air comes out cold; the design is exploiting regularities of the environment, rules of the world, laws of physics, that they don't know about. The domain of human thought and human brains is very poorly understood by us, and exhibits phenomena like optical illusions, hypnosis, psychosis, mania, or simple afterimages produced by strong stimuli in one place leaving neural effects in another place. Maybe a superintelligence couldn't defeat a human in a very simple realm like logical tic-tac-toe; if you're fighting it in an incredibly complicated domain you understand poorly, like human minds, you should expect to be defeated by 'magic' in the sense that even if you saw its strategy you would not understand why that strategy worked. **AI-boxing can only work on relatively weak AGIs; the human operators are not secure systems**.

Section C:

Okay, those are some significant problems, but lots of progress is being made on solving them, right? There's a whole field calling itself "AI Safety" and many major organizations are expressing Very Grave Concern about how "safe" and "ethical" they are?

37. There's a pattern that's played out quite often, over all the times the Earth has spun around the Sun, in which some bright-eyed young scientist, young engineer, young entrepreneur, proceeds in full bright-eyed optimism to challenge some problem that turns out to be really quite difficult. Very often the cynical old veterans of the field try to warn them about this, and the bright-eyed youngsters don't listen, because, like, who wants to hear about all that stuff, they want to go solve the problem! Then this person gets beaten about the head with a slipper by reality as they find out that their brilliant speculative theory is wrong, it's actually really hard to build the thing because it keeps breaking, and society isn't as eager to adopt their clever innovation as they might've hoped, in a process which eventually produces a new cynical old veteran. Which, if not literally optimal, is I suppose a nice life cycle to nod along to in a nature-show sort of way. Sometimes you do something for the *first* time and there *are* no cynical old veterans to warn anyone and people can be *really* optimistic about how it will go; eg the initial Dartmouth Summer Research Project on Artificial Intelligence in 1956: "An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer." This is *less* of a viable survival plan for your *planet* if the first major failure of the bright-eyed youngsters kills *literally everyone* before they can predictably get beaten about the head with the news that there were all sorts of unforeseen difficulties and reasons why things were hard. You don't get any cynical old veterans, in this case, because everybody on Earth is dead. Once you start to suspect you're in that situation, you have to do the Bayesian thing and update now to the view you will predictably update to later: realize you're in a situation of being that bright-eyed person who is going to encounter Unexpected Difficulties later and end up a cynical old veteran - or would be, except for the part where you'll be dead along with everyone else. And become that cynical old veteran *right away,* before reality whaps you upside the head in the form of everybody dying and you not getting to learn. **Everyone else seems to feel that, so long as reality hasn't whapped them upside the head yet and smacked them down with the actual difficulties, they're free to go on living out the standard life-cycle and play out their role in the script and go on being bright-eyed youngsters; there's no cynical old veterans to warn them otherwise, after all, and there's no proof that everything won't go beautifully easy and fine, *given their bright-eyed total ignorance of what those later difficulties could be.***

38. **It does not appear to me that the field of 'AI safety' is currently being remotely productive on tackling its enormous lethal problems.** These problems are in fact out of reach; the contemporary field of AI safety has been selected to contain people who go to work in that field anyways. Almost all of them are there to tackle problems on which they can appear to succeed and publish a paper claiming success; if they can do that and get funded, why would they embark on a much more unpleasant project of trying something harder that they'll fail at, just so the human species can die with marginally more dignity? This field is not making real progress and does not have a recognition function to distinguish real progress if it took place. You could pump a billion dollars into it and it would produce mostly noise to drown out what little progress was being made elsewhere.

39. **I figured this stuff out using the null string as input,** and frankly, I have a hard time myself feeling hopeful about getting real alignment work out of somebody who previously sat around waiting for somebody else to input a persuasive argument into them. This ability to "notice lethal difficulties without Eliezer Yudkowsky arguing you into noticing them" currently is an opaque piece of cognitive machinery to me, I do not know how to train it into others. It probably relates to 'security mindset', and a mental motion where you refuse to play out scripts, and being able to operate in a field that's in a state of chaos.

40. "Geniuses" with nice legible accomplishments in fields with tight feedback loops where it's easy to determine which results are good or bad right away, and so validate that this person is a genius, are (a) people who might not be able to do equally great work away from tight feedback loops, (b) people who chose a field where their genius would be nicely legible even if that maybe wasn't the place where humanity most needed a genius, and (c) probably don't have the mysterious gears simply because they're *rare.* **You cannot just pay $5 million apiece to a bunch of legible geniuses from other fields and expect to get great alignment work out of them.** They probably do not know where the real difficulties are, they probably do not understand what needs to be done, *they cannot tell the difference between good and bad work*, and the funders also can't tell without me standing over their shoulders evaluating everything, which I do not have the physical stamina to do. I concede that real high-powered talents, especially if they're still in their 20s, genuinely interested, and have done their reading, are people who, yeah, fine, have higher probabilities of making core contributions than a random bloke off the street. But I'd have more hope - not significant hope, but *more* hope - in separating the concerns of (a) credibly promising to pay big money retrospectively for good work to anyone who produces it, and (b) venturing prospective payments to somebody who is predicted to maybe produce good work later.

41. **Reading this document cannot make somebody a core alignment researcher**. That requires, not the ability to read this document and nod along with it, but the ability to spontaneously write it from scratch without anybody else prompting you; that is what makes somebody a peer of its author. It's guaranteed that some of my analysis is mistaken, though not necessarily in a hopeful direction. The ability to do new basic work noticing and fixing those flaws is the same ability as the ability to write this document before I published it, which nobody apparently did, despite my having had other things to do than write this up for the last five years or so. Some of that silence may, possibly, optimistically, be due to nobody else in this field having the ability to write things comprehensibly - such that somebody out there had the knowledge to write all of this themselves, if they could only have written it up, but they couldn't write, so didn't try. I'm not particularly hopeful of this turning out to be true in real life, but I suppose it's one possible place for a "positive model violation" (miracle). The fact that, twenty-one years into my entering this death game, seven years into other EAs noticing the death game, and two years into even normies starting to notice the death game, it is still Eliezer Yudkowsky writing up this list, says that humanity still has only one gamepiece that can do that. I knew I did not actually have the physical stamina to be a star researcher, I tried really really hard to replace myself before my health deteriorated further, and yet here I am writing this. That's not what surviving worlds look like.

42. **There's no plan.** Surviving worlds, by this point, and in fact several decades earlier, have a plan for how to survive. It is a written plan. The plan is not secret. In this non-surviving world, there are no candidate plans that do not immediately fall to Eliezer instantly pointing at the giant visible gaping holes in that plan. Or if you don't know who Eliezer is, you don't even realize you need a plan, because, like, how would a human being possibly realize that without Eliezer yelling at them? It's not like people will yell at *themselves* about prospective alignment difficulties, they don't have an *internal* voice of caution. So most organizations don't have plans, because I haven't taken the time to personally yell at them. 'Maybe we should have a plan' is deeper alignment mindset than they possess without me standing constantly on their shoulder as their personal angel pleading them into... continued noncompliance, in fact. Relatively few are aware even that they should, to look better, produce a *pretend* plan that can fool EAs too 'modest' to trust their own judgments about seemingly gaping holes in what serious-looking people apparently believe.

43. **This situation you see when you look around you is not what a surviving world looks like.** The worlds of humanity that survive have plans. They are not leaving to one tired guy with health problems the entire responsibility of pointing out real and lethal problems proactively. Key people are taking internal and real responsibility for finding flaws in their own plans, instead of considering it their job to propose solutions and somebody else's job to prove those solutions wrong. That world started trying to solve their important lethal problems earlier than this. Half the people going into string theory shifted into AI alignment instead and made real progress there. When people suggest a planetarily-lethal problem that might materialize later - there's a lot of people suggesting those, in the worlds destined to live, and they don't have a special status in the field, it's just what normal geniuses there do - they're met with either solution plans or a reason why that shouldn't happen, not an uncomfortable shrug and 'How can you be sure that will happen' / 'There's no way you could be sure of that now, we'll have to wait on experimental evidence.'

A lot of those better worlds will die anyways. It's a genuinely difficult problem, to solve something like that on your first try. But they'll die with more dignity than this.

A central AI alignment problem: capabilities generalization, and the sharp left turn

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(*This post was factored out of a larger post that I (Nate Soares) wrote, with help from Rob Bensinger, who also rearranged some pieces and added some text to smooth things out. I'm not terribly happy with it, but am posting it anyway (or, well, having Rob post it on my behalf while I travel) on the theory that it's better than nothing.*)

I expect navigating the acute risk period to be tricky for our civilization, for a number of reasons. Success looks to me to require clearing a variety of technical, sociopolitical, and moral hurdles, and while in principle sufficient mastery of solutions to the technical problems might substitute for solutions to the sociopolitical and other problems, it nevertheless looks to me like we need a lot of things to go right.

Some sub-problems look harder to me than others. For instance, people are still regularly surprised when I tell them that I think the hard bits are much more technical than moral: it looks to me like figuring out how to aim an AGI at all is harder than figuring out where to aim it.[1]

Within the list of technical obstacles, there are some that strike me as more central than others, like "figure out how to aim optimization". And a big reason why I'm currently fairly pessimistic about humanity's odds is that it seems to me like almost nobody is focusing on the technical challenges that seem most central and unavoidable to me.

Many people wrongly believe that I'm pessimistic because I think the alignment problem is extraordinarily difficult on a purely technical level. That's flatly false, and is pretty high up there on my list of least favorite misconceptions of my views.[2]

I think the problem is a normal problem of mastering some scientific field, as humanity has done many times before. Maybe it's somewhat trickier, on account of (e.g.) intelligence being more complicated than, say, physics; maybe it's somewhat easier on account of how we have more introspective access to a working mind than we have to the low-level physical fields; but on the whole, I doubt it's all that qualitatively different than the sorts of summits humanity has surmounted before.

It's made trickier by the fact that we probably have to attain mastery of general intelligence before we spend a bunch of time working with general intelligences (on account of how we seem likely to kill ourselves by accident within a few years, once we have AGIs on hand, if no pivotal act occurs), but that alone is not enough to undermine my hope.

What undermines my hope is that nobody seems to be working on the hard bits, and I don't currently expect most people to become convinced that they need to solve those hard bits until it's too late.

Below, I'll attempt to sketch out what I mean by "the hard bits" of the alignment problem. Although these look hard, I’m a believer in the capacity of humanity to solve technical problems at this level of difficulty when we put our minds to it. My concern is that I currently don’t think the field is *trying* to solve this problem. My hope in writing this post is to better point at the problem, with a follow-on hope that this causes new researchers entering the field to attack what seem to me to be the central challenges head-on.

Discussion of a problem

On my model, one of the most central technical challenges of alignment—and one that every viable alignment plan will probably need to grapple with—is the issue that capabilities generalize better than alignment.

My guess for how AI progress goes is that at some point, some team gets an AI that starts generalizing sufficiently well, sufficiently far outside of its training distribution, that it can gain mastery of fields like physics, bioengineering, and psychology, to a high enough degree that it more-or-less singlehandedly threatens the entire world. Probably without needing explicit training for its most skilled feats, any more than humans needed many generations of killing off the least-successful rocket engineers to refine our brains towards rocket-engineering before humanity managed to achieve a moon landing.

And in the same stroke that its capabilities leap forward, its alignment properties are revealed to be shallow, and to fail to generalize. The central analogy here is that optimizing apes for inclusive genetic fitness (IGF) doesn't make the resulting humans optimize mentally for IGF. Like, sure, the apes are eating because they have a hunger instinct and having sex because it feels good—but it's not like they *could* be eating/fornicating due to explicit reasoning about how those activities lead to more IGF. They can't yet perform the sort of abstract reasoning that would correctly justify those actions in terms of IGF. And then, when they start to generalize well in the way of humans, they predictably don't *suddenly start* eating/fornicating *because* of abstract reasoning about IGF, even though they now *could*. Instead, they invent condoms, and fight you if you try to remove their enjoyment of good food (telling them to just calculate IGF manually). The alignment properties you lauded before the capabilities started to generalize, predictably fail to generalize with the capabilities.

Some people I say this to respond with arguments like: "Surely, before a smaller team could get an AGI that can master subjects like biotech and engineering well enough to kill all humans, some other, larger entity such as a state actor will have a somewhat worse AI that can handle biotech and engineering somewhat less well, but in a way that prevents any one AGI from running away with the whole future?"

I respond with arguments like, " In the one real example of intelligence being developed we have to look at, continuous application of natural selection in fact found *Homo sapiens sapiens* , and the capability-gain curves of the ecosystem for various measurables were in fact sharply kinked by this new species (e.g., using machines, we sharply outperform other animals on well-established metrics such as “airspeed”, “altitude”, and “cargo carrying capacity”). "

Their response in turn is generally some variant of "well, natural selection wasn't optimizing very intelligently" or "maybe humans weren't all that sharply above evolutionary trends" or "maybe the power that let humans beat the rest of the ecosystem was simply the invention of culture, and nothing embedded in our own already-existing culture can beat us" or suchlike.

Rather than arguing further here, I'll just say that failing to believe the hard problem exists is one surefire way to avoid tackling it.

So, flatly summarizing my point instead of arguing for it: it looks to me like there will at some point be some sort of "sharp left turn", as systems start to work really well in domains really far beyond the environments of their training—domains that allow for significant reshaping of the world, in the way that humans reshape the world and chimps don't. And that's where (according to me) things start to get crazy. In particular, I think that once AI capabilities start to generalize in this particular way, it’s predictably the case that the alignment of the system will fail to generalize with it.[3]

This is slightly upstream of a couple other challenges I consider quite core and difficult to avoid, including:

1. Directing a capable AGI towards an objective of your choosing.
2. Ensuring that the AGI is low-impact, conservative, shutdownable, and otherwise corrigible.

These two problems appear in the strawberry problem , which Eliezer's been pointing at for quite some time: the problem of getting an AI to place two identical (down to the cellular but not molecular level) strawberries on a plate, and then do nothing else. The demand of cellular-level copying forces the AI to be capable; the fact that we can get it to duplicate a strawberry instead of doing some other thing demonstrates our ability to direct it; the fact that it does nothing else indicates that it's corrigible (or really well aligned to a delicate human intuitive notion of inaction).

How is the "capabilities generalize further than alignment" problem upstream of these problems? Suppose that the fictional team OpenMind is training up a variety of AI systems, before one of them takes that sharp left turn. Suppose they've put the AI in lots of different video-game and simulated environments, and they've had good luck training it to pursue an objective that the operators described in English. "I don't know what those MIRI folks were talking about; these systems are easy to direct; simple training suffices", they say. At the same time, they apply various training methods, some simple and some clever, to cause the system to allow itself to be removed from various games by certain "operator-designated" characters in those games, in the name of shutdownability. And they use various techniques to prevent it from stripmining in Minecraft, in the name of low-impact. And they train it on a variety of moral dilemmas, and find that it can be trained to give correct answers to moral questions (such as "in thus-and-such a circumstance, should you poison the operator's opponent?") just as well as it can be trained to give correct answers to any other sort of question. "Well," they say, "this alignment thing sure was easy. I guess we lucked out."

Then, the system takes that sharp left turn,[4][5] and, predictably, the capabilities quickly improve outside of its training distribution, while the alignment falls apart.

The techniques OpenMind used to train it away from the error where it convinces itself that bad situations are unlikely? Those generalize fine. The techniques you used to train it to allow the operators to shut it down? Those fall apart, and the AGI starts wanting to avoid shutdown, including wanting to deceive you if it’s useful to do so.

Why does alignment fail while capabilities generalize, at least by default and in predictable practice? In large part, because good capabilities form something like an attractor well. (That's one of the reasons to expect intelligent systems to eventually make that sharp left turn if you push them far enough, and it's why natural selection managed to stumble into general intelligence with no understanding, foresight, or steering.)

Many different training scenarios are teaching your AI the same instrumental lessons, about how to think in accurate and useful ways. Furthermore, those lessons are underwritten by a simple logical structure, much like the simple laws of arithmetic that abstractly underwrite a wide variety of empirical arithmetical facts about what happens when you add four people's bags of apples together on a table and then divide the contents among two people.

But that attractor well? It's got a free parameter. And that parameter is what the AGI is optimizing for. And there's no analogously-strong attractor well pulling the AGI's objectives towards your preferred objectives.

The hard left turn? That's your system sliding into the capabilities well. (You don't need to fall all that far to do impressive stuff; humans are better at an enormous variety of relevant skills than chimps, but they aren't all that lawful in an absolute sense.)

There's no analogous alignment well to slide into.

On the contrary, sliding down the capabilities well is liable to break a bunch of your existing alignment properties.[6]

Why? Because things in the capabilities well have instrumental incentives that cut against your alignment patches. Just like how your previous arithmetic errors (such as the pebble sorters on the wrong side of the Great War of 1957) get steamrolled by the development of arithmetic, so too will your attempts to make the AGI low-impact and shutdownable ultimately (by default, and in the absence of technical solutions to core alignment problems) get steamrolled by a system that pits those reflexes / intuitions / much-more-alien-behavioral-patterns against the convergent instrumental incentive to survive the day.

Perhaps this is not convincing; perhaps to convince you we'd need to go deeper into the weeds of the various counterarguments, if you are to be convinced. (Like acknowledging that humans, who can foresee these difficulties and adjust their training procedures accordingly, have a *better* chance than natural selection did, while then discussing why current proposals do not seem to me to be hopeful.) But hopefully you can at least, in reading this document, develop a basic understanding of my position.

Stating it again, in summary: my position is that capabilities generalize further than alignment (once capabilities start to generalize real well (which is a thing I predict will happen)). And this, by default, ruins your ability to direct the AGI (that has slipped down the capabilities well), and breaks whatever constraints you were hoping would keep it corrigible. And addressing the problem looks like finding some way to either keep your system aligned through that sharp left turn, or render it aligned afterwards.

In an upcoming post (**edit**: here), I’ll say more about how it looks to me like ~nobody is working on this particular hard problem, by briefly reviewing a variety of current alignment research proposals. In short, I think that the field’s current range of approaches nearly all assume this problem away, or direct their attention elsewhere.

1. ^

   Furthermore, figuring where to aim it looks to me like more of a technical problem than a moral problem. Attempting to manually specify the nature of goodness is a doomed endeavor, of course, but that's fine, because we can instead specify processes for figuring out (the coherent extrapolation of) what humans value. Which still looks prohibitively difficult as a goal to give humanity's first AGI (which I expect to be deployed under significant time pressure), mind you, and I further recommend aiming humanity's first AGI systems at simple limited goals that end the acute risk period and then cede stewardship of the future to some process that can reliably do the "aim minds towards the right thing" thing. So today's alignment problems are a few steps removed from tricky moral questions, on my models.

2. ^

   While we're at it: I think trying to get provable safety guarantees about our AGI systems is silly, and I'm pretty happy to follow Eliezer in calling an AGI "safe" if it has a <50% chance of killing >1B people. Also, I think there's a very large chance of AGI killing us, and I thoroughly disclaim the argument that even if the probability is tiny then we should work on it anyway because the stakes are high.

3. ^

   Note that this is consistent with findings like “large language models perform just as well on moral dilemmas as they perform on non-moral ones”; to find this reassuring is to misunderstand the problem. Chimps have an easier time than squirrels following and learning from human cues. Yet this fact doesn't particularly mean that enhanced chimps are more likely than enhanced squirrels to remove their hunger drives, once they understand inclusive genetic fitness and are able to eat purely for reasons of fitness maximization. Pre-left-turn AIs will get better at various 'alignment' metrics, in ways that I expect to build a false sense of security, without addressing the lurking difficulties.

4. ^

   "What do you mean ‘it takes a sharp left turn’? Are you talking about recursive self-improvement? I thought you said somewhere else that you don't think recursive self-improvement is necessarily going to play a central role before the extinction of humanity?" I'm not talking about recursive self-improvement. That's one way to take a sharp left turn, and it could happen, but note that humans have neither the understanding nor control over their own minds to recursively self-improve, and we outstrip the rest of the animals pretty handily. I'm talking about something more like “intelligence that is general enough to be dangerous”, the sort of thing that humans have and chimps don't.

5. ^

"Hold on, isn't this unfalsifiable? Aren't you saying that you're going to continue believing that alignment is hard, even as we get evidence that it's easy?" Well, I contend that "GPT can learn to answer moral questions just as well as it can learn to answer other questions" is not much evidence either way about the difficulty of alignment. I'm not saying we'll get evidence that I'll ignore; I'm naming in advance some things that I wouldn't consider negative evidence (partially in hopes that I can refer back to this post when people crow later and request an update). But, yes, my model does have the inconvenient property that people who are skeptical now, are liable to remain skeptical until it's too late, because most of the evidence I expect to give us *advance* warning about the nature of the problem is evidence that we've already seen. I assure you that I do not consider this property to be convenient.

As for things that could convince me otherwise: technical understanding of intelligence could undermine my "sharp left turn" model. I could also imagine observing some ephemeral hopefully-I'll-know-it-when-I-see-it capabilities thresholds, without any sharp left turns, that might update me. (Short of "full superintelligence without a sharp left turn", which would obviously convince me but comes too late in the game to shift my attention.)

6. ^

   To use my overly-detailed evocative example from earlier: Humans aren't tempted to rewire our own brains so that we stop liking good meals for the sake of good meals, and start eating only insofar as we know we have to eat to reproduce (or, rather, maximize inclusive genetic fitness) (after upgrading the rest of our minds such that that sort of calculation doesn't drag down the rest of the fitness maximization). The cleverer humans are chomping at the bit to have their beliefs be more accurate, but they're not chomping at the bit to replace all these mere-shallow-correlates of inclusive genetic fitness with explicit maximization. So too with other minds, at least by default: that which makes them generally intelligent, does not make them motivated by your objectives.

On how various plans miss the hard bits of the alignment challenge

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This post has been recorded as part of the LessWrong Curated Podcast, and can be listened to on Spotify , Apple Podcasts , and Libsyn .

(As usual, this post was written by Nate Soares with some help and editing from Rob Bensinger.)

In my last post, I described a “hard bit” of the challenge of aligning AGI—the sharp left turn that comes when your system slides into the “AGI” capabilities well, the fact that alignment doesn’t generalize similarly well at this turn, and the fact that this turn seems likely to break a bunch of your existing alignment properties.

Here, I want to briefly discuss a variety of current research proposals in the field, to explain why I think this problem is currently neglected.

I also want to mention research proposals that *do* strike me as having some promise, or that strike me as adjacent to promising approaches.

Before getting into that, let me be very explicit about three points:

1. On my model, solutions to how capabilities generalize further than alignment are necessary but not sufficient. There is dignity in attacking a variety of other real problems, and I endorse that practice.
2. The imaginary versions of people in the dialogs below are not the same as the people themselves. I'm probably misunderstanding the various proposals in important ways, and/or rounding them to stupider versions of themselves along some important dimensions.[1] If I've misrepresented your view, I apologize.
3. I do not subscribe to the Copenhagen interpretation of ethics wherein someone who takes a bad swing at the problem (or takes a swing at a different problem) is more culpable for civilization's failure than someone who never takes a swing at all. Everyone whose plans I discuss below is highly commendable, laudable, and virtuous by my accounting.

Also, many of the plans I touch upon below are not being given the depth of response that I'd ideally be able to give them, and I apologize for not engaging with their authors in significantly more depth first. I’ll be especially cursory in my discussion of some MIRI researchers and research associates like Vanessa Kosoy and Scott Garrabrant.[2]

In this document I'm attempting to summarize my high-level view of the approaches I know about; I'm not attempting to provide full arguments for why I think particular approaches are more or less promising.

Think of the below as a window into my thought process, rather than an attempt to state or justify my entire background view. And obviously, if you disagree with my thoughts, I welcome objections.

So, without further ado, I'll explain why I think that the larger field is basically not working on this particular hard problem:

Reactions to specific plans

Owen Cotton-Barratt & Truthful AI

Imaginary, possibly-mischaracterized-by-Nate version of Owen: What if we train our AGIs to be truthful? If our AGIs were generally truthful, we could just ask them if they're plotting to be deceptive, and if so how to fix it, and we could do these things early in ways that help us nip the problems in the bud before they fester, and so on and so forth.

Even if that particular idea doesn't work, it seems like our lives are a lot easier insofar as the AGI is truthful.

Nate: "Truthfulness" sure does sound like a nice property for our AGIs to have. But how do you get it in there? And how do you *keep* it in there, after that sharp left turn? If this idea is to make any progress on the hard problem we're discussing, it would have to come from some property of "truthfulness" that makes it more likely than other desirable properties to survive the great generalization of capabilities.

Like, even simpler than the problem of an AGI that puts two identical strawberries on a plate and does nothing else , is the problem of an AGI that turns as much of the universe as possible into diamonds. This is easier because, while it still requires that we have some way to direct the system towards a concept of our choosing, we no longer require corrigibility. (Also, "diamond" is a significantly simpler concept than "strawberry" and "cellularly identical".)

It seems to me that we have basically no idea how to do this. We can train the AGI to be pretty good at building diamond-like things across a lot of training environments, but once it takes that sharp left turn, *by default* , it will wander off and do some other thing, like how humans wandered off and invented birth control .

In my book, solving this hard problem so well that we could feasibly get an AGI that predictably maximizes diamond (after its capabilities start generalizing hard), would constitute an enormous advance.

Solving the hard problem so well that we could feasibly get an AGI that predictably *answers operator questions truthfully*, would constitute a similarly enormous advance. Because we would have figured out how to keep a highly capable system directed at any one thing of our choosing.

Now, in real life, building a truthful AGI is much harder than building a diamond optimizer, because 'truth' is a concept that's much more fraught than 'diamond'. (To see this, observe that the definition of "truth" routes through tricky concepts like "ways the AI communicated with the operators" and "the mental state of the operators", and involves grappling with tricky questions like "what ways of translating the AI's foreign concepts into human concepts count as manipulative?" and "what can be honestly elided?", and so on, whereas diamond is just carbon atoms bound covalently in tetrahedral lattices.)

So as far as I can tell, from the perspective of this hard problem, Owen's proposal boils down to "Wouldn't it be nice if the tricky problems were solved, and we managed to successfully direct our AGIs to be truthful?" Well, sure, that would be nice, but it's not helping solve our problem. In fact, this problem subsumes the whole diamond maximizer problem, but replaces the concept of "diamond" (that we obviously can't yet direct an AGI to optimize, diamond *more clearly* being a physical phenomenon far removed from the AGI's raw sensory inputs) with the concept of "truth" (which is abstract enough that we can easily forget that it's a *much more difficult-to-describe* physical phenomenon far removed from the AGI's raw sensory inputs).

(And this hasn't even touched on how if you *did* manage to get an AGI that keeps optimizing for truth-telling after it falls into the capabilities well, then you still have to make it corrigible, on pain of extinction. But I digress.)

Maybe Owen does think that the goal of "tell the truth" generalizes more readily than "optimize diamond", for some reason? But if so, I haven't seen him mention it, except when I explicitly prompted him when having this argument in person. Which doesn't seem like a very promising sign to me. My read is that he's just failing to engage with this hard problem.

Ryan Greenblatt & Eliciting Latent Knowledge

Imaginary Ryan, because Paul was on vacation last time I tried to understand ELK: As we train the nascent AI up to an AGI, we're going to simultaneously train a second system (the "ELK head") that can elicit its latent knowledge. The latter system will be able to tell us the moment that the former system is starting to do things like contemplate deception. We're then going to combine this with some third adversarial system, which can use that latent knowledge to expose the bad behavior to gradients, so that we can nip it in the bud.

Basically all this requires is that, whatever intelligence gradient descent (or some other optimizing procedure) can create, it can also decode, which doesn't seem all that big an ask: whatever optimization method you were using that found this intelligent reasoner, there should be a dual optimization process that finds its decoder, if you're a little clever about it.

Nate: So, set aside questions of whether that sort of thing (with the separate knowledge-elicitor and adversary) will turn out to be practical, because we're talking about whether the plan attacks the key problem.

Suppose we have an ELK head, and an adversary, that we're using to train a nascent AGI. The AGI takes a sharp left turn. Now it's dangerously capable, and suddenly a bunch of misalignments which it previously didn't notice within itself are relevant. And you've got to get them out.

Your first problem is that the recent capabilities gains made by the AGI might not have come from gradient descent (much like how humans’ sudden explosion of technological knowledge accumulated in our culture rather than our genes, once we turned the corner). You might not be able to just "expose the bad behavior" to gradients that you can hit to correct the thing, at least not easily and quickly.

Your second problem is that the AGI's concepts might rapidly get totally uninterpretable to your ELK head. Like, you could imagine doing neuroimaging on your mammals all the way through the evolution process. They've got some hunger instincts in there, but it's not like they’re smart enough yet to represent the concept of "inclusive genetic fitness" correctly, so you figure you'll just fix it when they get capable enough to understand the alternative (of eating because it's instrumentally useful for procreation). And so far you're doing great: you've basically decoded the visual cortex, and have a pretty decent understanding of what it's visualizing.

You maintain this pretty well as it walks up through to primate, and then suddenly it takes a sharp left turn and invents its own internal language and a bunch of abstract concepts, and suddenly you find your visualization tools to be quite lacking for interpreting its abstract mathematical reasoning about topology or whatever.

Analogously, your ELK head's abilities are liable to fall off a cliff right as the AGI's capabilities start generalizing way outside of its training distribution.

And if they don't, then this ELK head is (in this hypothetical) able to decode and understand the workings of an alien mind. Likely a kludgey behemoth of an alien mind. This itself is liable to require quite a lot of capability, quite plausibly of the sort that humanity gets first from the systems that took sharp left-turns, rather than systems that ground along today's scaling curves until they scaled that far.

Or in other words, if your ELK head does keep pace with your AGI, and takes a sharp left turn at the same time as it... then, well, now you're basically back to the "Truthful AI" proposal. How do you keep your ELK head reporting accurately (and doing so corrigibly), as it undergoes that sharp left turn?

This proposal seems to me like it's implicitly assuming that most of the capabilities gains come from the slow grind of gradient descent, in a world where the systems don't take sharp left turns and rapidly become highly capable in a wide variety of new (out-of-distribution) domains.

Which seems to me that it's mostly just assuming its way out from under the hard problem—and thus, on my models, assuming its way clean out of reality.

And if I imagine attempting to apply this plan inside of the reality I think I live in, I don't see how it plans to address the hard part of the problem, beyond saying "try training it against places where it knows it's diverging from the goal before the sharp turn, and then hope that it generalizes well or won't fight back", which doesn't instill a bunch of confidence in me (and which I don't expect to work).

Eric Drexler & AI Services

Imaginary Eric: Well, sure, AGI could get real dangerous if you let one system do everything under one umbrella. But that's not how good engineers engineer things. You can and should split your AI systems into siloed services, each of which can usefully help humanity with some fragment of whichever difficult sociopolitical or physical challenge you're hoping to tackle, but none of which constitutes an adversarial optimizer (with goals over the future) in its own right.

Nate: So mostly I expect that, if you try to split these systems into services, then you either fail to capture the heart of intelligence and your siloed AIs are irrelevant, or you wind up with enough AGI in one of your siloes that you have a whole alignment problem (hard parts and all) in there.

Like, I see this plan as basically saying "yep, that hard problem is in fact too hard, let's try to dodge it, by having humans + narrow AI services perform the pivotal act". Setting aside how I don't particularly expect this to work, we can at least hopefully agree that it's attempting to route around the problems that seem to me to be central, rather than attempting to solve them.

Evan Hubinger, in a recent personal conversation

Imaginary Evan: It's hard, in the modern paradigm, to separate the system's values from its capabilities and from the way it was trained. All we need to do is find a training regimen that leads to AIs that are both capable and aligned. At which point we can just make it publicly available, because it's not like people will be trying to disalign their AIs.

Nate: So, first of all, you haven't exactly made the problem *easier*.

As best I can tell, this plan amounts to "find a training method that not only *can* keep a system aligned through the sharp left turn, but *must*, and then popularize it". Which has, like, bolted two additional steps atop an assumed solution to some hard problems. So this proposal does not seem, to me, to make any progress towards solving those hard problems.

(Also, the observation "capabilities and alignment are fairly tightly coupled in the modern paradigm" doesn't seem to me like much of an argument that they're going to *stay* coupled after the ol' left turn. Indeed, I expect they won't stay coupled in the ways you want them to. Assuming that this modern desirable property will hold indefinitely seems dangerously close to just assuming this hard problem away, and thus assuming your way clean out of what-I-believe-to-be-reality.)

But maybe I just don't understand this proposal yet (and I have had some trouble distilling things I recognize as plans out of Evan's writing, so far).

A fairly straw version of someone with technical intuitions like Richard Ngo's or

Rohin Shah’s

Imaginary Richard/Rohin: You seem awfully confident in this sharp left turn thing. And that the goals it was trained for *won't* just generalize. This seems characteristically overconfident. For instance, observe that natural selection didn't try to get the inner optimizer to be aligned with inclusive genetic fitness *at all*. For all we know, a small amount of cleverness in exposing inner-misaligned behavior to the gradients will just be enough to fix the problem. And even if not that-exact-thing, then there are all sorts of ways that some other thing could come out of left field and just render the problem easy. So I don't see why you're worried.

Nate: My model says that the hard problem rears its ugly head by default, in a pretty robust way. Clever ideas might suffice to subvert the hard problem (though my guess is that we need something more like understanding and mastery, rather than just a few clever ideas). I have considered an array of clever ideas that look to me like they would predictably-to-me fail to solve the problems, and I admit that my guess is that you're putting most of your hope on small clever ideas that I can already see would fail. But perhaps you have ideas that I do not. Do you yourself have any specific ideas for tackling the hard problem?

Imaginary Richard/Rohin: Train it, while being aware of inner alignment issues, and hope for the best.

Nate: That doesn't seem to me to even start to engage with the issue where the capabilities fall into an attractor and the alignment doesn't.

Perhaps sometime we can both make a list of ways to train with inner alignment issues in mind, and then share them with each other, so that you can see whether you think I'm lacking awareness of some important tool you expect to be at our disposal, and so that I can go down your list and rattle off the reasons why the proposed training tools don't look to me like they result in alignment that is robust to sharp left turns. (Or find one that surprises me, and update.) But I don't want to delay this post any longer, so, some other time, maybe.

Another recent proposal

Imaginary Anonymous-Person-Whose-Name-I’ve-Misplaced: Okay, but maybe there is a pretty wide attractor basin around my own values, though. Like, maybe not my true values, but around a bunch of stuff like being low-impact and deferring to the operators about what to do and so on. You don't need to be all that smart, nor have a particularly detailed understanding of the subtleties of ethics, to figure out that it's bad (according to me) to kill all humans.

Nate: Yeah, that's basically the idea behind corrigibility, and is one reason why corrigibility is plausibly a lot easier to get than a full-fledged CEV sovereign. But this observation doesn't really engage with the question of *how to point the AGI towards that concept*, and *how to cause its behavior to be governed by that concept* in a fashion that's robust to the sharp left turn where capabilities start to really generalize.

Like, yes, some directions are easier to point an AI in, on account of the direction itself being simpler to conceptualize, but that observation alone doesn't say anything about how to determine which direction an AI is pointing after it falls into the capabilities well.

More generally, saying "maybe it's easy" is not the same as solving the problem. Maybe it is easy! But it's not going to get solved unless we have people trying to solve it.

Vivek Hebbar, summarized (perhaps poorly) from last time we spoke of this in person

Imaginary Vivek: Hold on, the AGI is being taught about what I value every time it tries something and gets a gradient about how well that promotes the thing I value. At least, assuming for the moment that we have a good ability to evaluate the goodness of the consequences of a given action (which seems fair, because it sounds like you're arguing for a way that we'd be screwed even if we had the One True Objective Function).

Like, you said that all aspects of reality are whispering to the nascent AGI of what it means to optimize, but few parts of reality are whispering of what to optimize for—whereas it looks to me like every gradient the AGI gets is whispering a little bit of both. So in particular, it seems to me like if you *did* have the one true objective function, you could just train good and hard until the system was both capable and aligned.

Nate: This seems to me like it's implicitly assuming that all of the system's cognitive gains come from the training. Like, with every gradient step, we are dragging the system one iota closer to being capable, and also one iota closer to being good, or something like that.

To which I say: I expect many of the cognitive gains to come from elsewhere, much as a huge number of the modern capabilities of humans are encoded in their culture and their textbooks rather than in their genomes. Because there are slopes in capabilities-space that an intelligence can snowball down, picking up lots of cognitive gains, but not alignment, along the way.

Assuming that this is not so, seems to me like simply assuming this hard problem away.

And maybe you simply don't believe that it's a real problem; that's fine, and I'd be interested to hear why you think that. But I have not yet heard a proposed *solution*, as opposed to an objection to the existence of the problem in the first place.

John Wentworth & Natural Abstractions

Imaginary John: I suspect there's a common format to concepts, that is a fairly objective fact about the math of the territory, and that—if mastered—could be used to understand an AGI's concepts. And perhaps select the ones we wish it would optimize for. Which isn't the whole problem, but sure is a big chunk of the problem. (And other chunks might well be easier to address given mastery of the fairly-objective concepts of "agent" and "optimizer" and so on.)

Nate: This does seem to me like it's trying to attack the actual problem! I have my doubts about this particular line of research (and those doubts are on my list of things to write up), but hooray for a proposal that, if it succeeded by its own lights, would address this hard problem!

Imaginary John: Well, uh, these days I'm mostly focusing on using my flimsy non-mastered grasp of the common-concept format to try to give a descriptive account of human values, because for some reason that's where I think the hope is. So I'm not *actually* working too much on this thing that you think takes a swing at the real problem (although I do flirt with it occasionally).

Nate: :'(

Imaginary John: Look, I didn't want to break the streak, OK.

Rob Bensinger, reading this draft: Wait, why do you see John’s proposal as attacking the central problem but not, for example, Eric Drexler’s Language for Intelligent Machines (summarized here)?

Nate: I understand Eric to be saying "maybe humans deploying narrow AIs will be capable enough to end the acute risk period before an AGI can (in which case we can avoid ever using AIs that have taken sharp left turns)", whereas John is saying "maybe a lot of objective facts about the territory determine which concepts are useful, and by understanding the objectivity of concepts we can become able to understand even an alien mind's concepts".

I think John’s guess is *wrong* (at least in the second clause), but it seems aimed at taking an AI system that has snowballed down a capabilities slope in the way that humans snowballed, and identifying its concepts in a way that’s stable to changes in the AI’s ontology—which is step one in the larger challenge of figuring out how to robustly direct an AGI’s motivations at the content of a particular concept it has.

My understanding of Eric’s idea, in contrast, is "I think there's a language these siloed components could use that's not so expressive as to allow them to be dangerous, but is expressive enough to allow them to help humans." To which my basic reply is roughly “The problem is that the non-siloed systems are going to start snowballing and end the world before the human+silo systems can save the world." As far as I can tell, Eric's attempting to route around the problem, whereas John's attempting to solve it.[3]

Neel Nanda & Theories of Impact for Interpretability

Imaginary Neel: What if we get a lot of interpretability?

Nate: That would be great, and I endorse developing such tools.

I think this will only solve the hard problems if the field succeeds at interpretability *so wildly* that (a) our interpretability tools continue to work on fairly difficult concepts in a post-left-turn AGI; (b) that AGI has an architecture that turns out to be especially amenable to being aimed at some concept of our choosing; and (c) the interpretability tools grant us such a deep understanding of this alien mind that we can aim it using that understanding.

I admit I'm skeptical of all three. Where, to be clear, better interpretability tools help put us in a better position even if they don't clear these lofty bars. In real life, I expect interpretability to play a smaller role as a force-multiplier that awaits some other plan for addressing the hard problems.

Which are great to have and worth building, to be clear. I full-throatedly endorse humanity putting more effort into interpretability.

It simultaneously doesn't look to me like people are seriously aiming for "develop such a good ability to understand minds that we can reshape/rebuild them to be aimable in whatever time we have after we get one". It looks to me like the sights are currently set at much lower and more achievable targets, and that current progress is consistent with never hitting the more ambitious targets, the ones that would let us understand and reshape the first artificial minds into something aligned (fast enough to be relevant).

But if some ambitious interpretability researchers do set their sights on the sharp left turn and the generalization problem, then I would indeed count this as a real effort by humanity to solve its central technical challenge. I don't need a lot of hope in a specific research program in order to be satisfied with the field's allocation of resources; I just want to grow the space of attempts to solve the generalization problem *at all*.

Stuart Armstrong & Concept Extrapolation

Nate: (*Note: This section consists of actual quotes and dialog, unlike the others.*)[4]

Stuart, in a blog post:

> […] It is easy to point at current examples of agents with low (or high) impact, at safe (or dangerous) suggestions, at low (or high) powered behaviours. So we have in a sense the 'training sets' for defining low-impact/Oracles/low-powered AIs.
>
> It's extending these examples to the general situation that fails: definitions which cleanly divide the training set (whether produced by algorithms or humans) fail to extend to the general situation. Call this the 'value extrapolation problem, with 'value' interpreted broadly as a categorisation of situations into desirable and undesirable.
>
> […] Value extrapolation is thus necessary for AI alignment.
>
> […] We think that once humanity builds its first AGI, superintelligence is likely near, leaving little time to develop AI safety at that point. Indeed, it may be necessary that the first AGI start off aligned: we may not have the time or resources to convince its developers to retrofit alignment to it. So we need a way

> to have alignment deployed throughout the algorithmic world before anyone develops AGI.
>
> To do this, we'll start by offering alignment as a service for more limited AIs. Value extrapolation scales down as well as up: companies value algorithms that won't immediately misbehave in new situations, algorithms that will become conservative and ask for guidance when facing ambiguity.
>
> We will get this service into widespread use (a process that may take some time), and gradually upgrade it to a full alignment process. [...]

Rob Bensinger, replying on [Twitter](#): The basic idea in that post seems to be: let's make it an industry standard for AI systems to "become conservative and ask for guidance when facing ambiguity", and gradually improve the standard from there as we figure out more alignment stuff.

The reasoning being something like: once we have AGI, we need to have deployment-ready aligned AGI *extremely soon*; and this will be more possible if the non-AGI preceding it is largely aligned.

(I at least agree with the "once we have AGI, we'll need deployment-ready aligned AGI extremely soon" part of this.)

The other aspect of your plan seems to be 'focus on improving value extrapolation methods'. Both aspects of this plan seem very bad to me, speaking from my inside view:

- 1a. I don't expect that much overlap between what's needed to make, e.g., a present-day image classifier more conservative, and what's needed to make an AGI reliable and safe. So redirecting resources from the latter problem to the former seems wasteful to me.
- 1b. Relatedly, I don't think it's helpful for the field to absorb the message "oh, yeah, our image classifiers and Go players and so on are aligned, we're knocking that problem out of the park". If 1a is right, then making your image classifier conservative doesn't represent much progress toward being able to align AGI. They're different problems, like building a safe bridge vs. building a safe elevator.

'Alignment' is currently a word that's about the AGI problem in particular, which overlaps with a lot of narrow-AI robustness problems, but isn't just a scaled-up version of those; the difficulty of AGI alignment mostly comes from qualitatively new risks. So 'aligning' the field as a whole doesn't necessarily help much, and (less importantly) using the *term* 'alignment' for the broader, fuzzier goal is liable to distract from the core difficulties, and liable to engender a false sense of progress on the original problem.

- 2. We need to do value extrapolation eventually, but I don't think this is the field's current big bottleneck, and I don't think it helps address the bottleneck. Rather, I think the big bottleneck is understandability / interpretability.

Nate: I like Rob's response. I'll add that I'm not sure I understand your proposal. Your previous name for the value extrapolation problem was the "model splintering" problem, and iirc you endorsed [Rohin's summary](#) of model splintering:

> [Model splintering] is one way of more formally looking at the out-of-distribution problem in machine learning: instead of simply saying that we are out of distribution, we look at the model that the AI previously had, and see what model it transitions to in the new distribution, and analyze this transition.
>
> Model splintering in particular refers to the phenomenon where a coarse-grained model is “splintered” into a more fine-grained model, with a one-to-many mapping between the environments that the coarse-grained model can distinguish between and the environments that the fine-grained model can distinguish between (this is what it means to be more fine-grained).

On the surface, work aimed at understanding and addressing "model splintering" sounds potentially promising to me—like, I might want to classify some version of "concept extrapolation" alongside Natural Abstractions, certain approaches to interpretability, Vanessa’s work, Scott’s work, etc. as "an angle of attack that might genuinely help with the core problem, if it succeeded wildly more than I expect it to succeed". Which is about as positive a situation as I’m expecting right now, and would be high praise in my books.

But in the past, I’ve often heard you use words and phrases in ways that I find promising at a glance, to mean things that I end up finding much less promising when I dig in on the specifics of what you’re talking about. So I’m initially skeptical, especially insofar as I don’t understand your proposal well.

I’d be interested in hearing how you think your proposal addresses the sharp left turn, if you think it does; or maybe you can give me pointers toward particular paragraphs/sections you’ve written up that you think already speak to this problem.

Regarding work on image-classifier conservatism: at a first glance, I don't have much confidence that the types of generalization you’re shooting for are tracking the possibility of sharp left turns. "We want our solutions to generalize" is cheap to say; things that engage with the sharp left turn are more expensive. What’s an example of a kind of present-day research on image classifier conservatism that you’d expect to help with the sharp left turn (if you do think any would help)?

Rebecca Gorman, in an email thread: We're working towards something that achieves interpretability objectives, and does so better than current approaches.

Agreed that AGI alignment isn't just a scaled-up version of narrow-AI robustness problems. But if we need to establish the foundations of alignment before we reach AGI and build it into every AI being built today (since we don't know when and where superintelligence will arise), then we need to try to scale *down* the alignment problem to something we can start to research today.

As for the article [A central AI alignment problem: capabilities generalization, and the sharp left turn]: I think it's an excellent article, but I'll give an insufficient response. I agree that capabilities form an attractor well. And that we don't get a strong understanding of human values as easily. That's why we think it's important to invest energy and resources into giving AI a strong understanding of human values; it's probably a harder problem. But - at a high level, some of the methods for getting there may generalize. That, at least, is a hopeful statement.

Nate: That sounds like a laudable goal. I have not yet managed to understand what sort of foundations of alignment you're trying to scale down and build into modern systems. What are you hoping to build into modern systems, and how do you expect it to relate to the problem of aligning systems with capabilities that generalize far outside of training?

So far, from parts of the aforementioned email thread that have been elided in this dialog, I have not yet managed to extract a plan beyond "generate training data that helps things like modern image classifiers distinguish intended features (such as ‘pre-treatment collapsed lung’ from ‘post-treatment collapsed lung with chest drains installed’, despite the chest-drains being easier to detect than the collapse itself)", and I don't yet see how generating this sort of training data and training modern image-classifiers thereon addresses the tricky alignment challenges I worry about.

Stuart, in an email thread: In simple or typical environments, simple proxies can achieve desired goals. Thus AIs tend to learn simple proxies, either directly (programmers write down what they currently think the goal is, leaving important pieces out) or indirectly (a simple proxy fits the training data they receive - eg image classifiers focusing on spurious correlations).

Then the AI develops a more complicated world model, either because the AI is becoming smarter or because the environment changes by itself. At this point, by the usual Goodhart arguments, the simple proxy no longer encodes desired goals, and can be actively pernicious.

What we're trying to do is to ensure that, when the AI transitions to a different world model, this updates its reward function at the same time. Capability increases should lead immediately to alignment increases (or at least alignment changes); this is the whole model splintering/value extrapolation approach.

The benchmark we published is a much-simplified example of this: the "typical environment" is the labeled datasets where facial expression and text are fully correlated. The "simple proxy/simple reward function" is the labeling of these images. The "more complicated world model" is the unlabeled data that the algorithm encounters, which includes images where the expression feature and the text feature are uncorrelated. The "alignment increase" (or, at least, the first step of this) is the algorithm realising that there are multiple distinct features in its "world model" (the unlabeled images) that could explain the labels, and thus generating multiple candidates for its "reward function".

One valid question worth asking is why we focused on image classification in a rather narrow toy example. The answer is that, after many years of work in this area, we've concluded that the key insights in extending reward functions do not lie in high-level philosophy, mathematics, or modelling. These have been useful, but have (temporarily?) run their course. Instead, practical experiments in value extrapolation seem necessary - and these will ultimately generate theoretical insights. Indeed, this has already happened; we now have, I believe, a much better understanding of model splintering than before we started working on this.

As a minor example, this approach seems to generate a new form of interpretability. When the algorithm asks the human to label a "smiling face with SAD written on it", it doesn't have a deep understanding of either expression or text; nor do humans have an understanding of what features it is really using. Nevertheless, seeing the ambiguous image gives us direct insight into the "reward functions" it is comparing, a potential new form of interpretability. There are other novel theoretical insights which we've been discussing in the company, but they're not yet written up for public presentation.

We're planning to generalise the approach and insights from image classifiers to other agent designs (RL agents, recommender systems, language models...); this will generate more insights and understanding on how value extrapolation works in general.

Nate: In Nate-speak, the main thing I took away from what you've said is "I want alignment to generalize when capabilities generalize. Also, we're hoping to get modern image classifiers to ask for labels on ambiguous data."

"Get the AI to ask for labels on ambiguous data" is one of many ideas I'd put on a list of shallow alignment ideas that are worth implementing. To my eye, it doesn't seem particularly related to the problem of pointing an AGI at something in a way that's robust to capabilities-start-generalizing.

It's a fine simple tool to use to help point at the concept you were hoping to point at, if you can get an AGI to do the thing you're pointing toward at all, and it would be embarrassing if we didn't try it. And I'm happy to have people trying early versions of such things as soon as possible. But I don't see these sorts of things as shedding much light on how you get a post-left-turn AGI to optimize for some concept of your choosing in the first place. If you could do that, then sure, getting it to ask for clarification when the training data is ambiguous is a nice extra saving throw (if it wasn't already doing that automatically because of some deeper corrigibility success), but I don't currently see this sort of thing as attacking one of the core issues.[5]

Andrew Critch & political solutions

Imaginary Andrew Critch: Just politick between the AGI teams and get them all to agree to take the problem seriously, not race, not cut corners on safety, etc.

Nate: Uh, that ship sailed in, like, late 2015. My fairly-strong impression, from my proximity to the current politics between the current orgs, is "nope".

Also, even if this wasn't a straight-up "nope", you have the question of what you *do* with your cooperation. Somehow you've still got to leverage this cooperation into the end of the acute risk period, before the people outside your alliance end the world. And this involves having a leadership structure that can distinguish bad plans from good ones.

The alliance helps, for sure. It takes a bunch of the time pressure off (assuming your management is legibly capable of distinguishing good deployment ideas from bad ones). I endorse attempts to form such an alliance. (And it sure would be undignified for our world to die of antitrust law at the final extremity.) But it's not an attempt to solve this hard technical problem, and it doesn't alleviate enough pressure to cause me to think that the problem would eventually be solved, in this field where ~nobody manages to strike for the heart of the problem before them.

Imaginary Andrew Critch: So get global coordination going! Or have some major nation-state regulate global use of AI, in some legitimate way!

Nate: Here I basically have the same response: First, can't be done (though I endorse attempts to prove me wrong, and recommend practicing by trying to effect important political change on smaller-stakes challenges ASAP (The time is ripe for sweeping global coordination in pandemic preparedness! We just had our warning shot! If we'll be able to do something about AGI later, presumably we can do something analogous about pandemics now!)).

Second, it doesn't alleviate *enough* pressure; the bureaucrats can't tell real solutions from bad ones; the cost to build an unaligned AGI drops each year; etc., etc. Sufficiently good global coordination is a win condition, but we're not anywhere close to on track for that, and in real life we're still going to need technical solutions.

Which, apparently, only a handful of people in the world are trying to provide.

What about superbabies?

Nate: I doubt we have the time, but sure, go for superbabies. It's as dignified as any of the other attempts to walk around this hard problem.

What about other MIRI people?

There are a few people supported at least in part by MIRI (such as Scott and Vanessa) who seem to me to have identified confusing and poorly-understood aspects of cognition. And their targets strike me as the sort of things where if we got less confused about what the heck was going on, then we might thereby achieve a somewhat better understanding of minds/optimization/etc., in a way that sheds some light on the hard problems. So yeah, I'd chalk a few other MIRI-supported folk up in the "trying to tackle the hard problems" column.

We still wouldn’t have anything close to a full understanding, and at the progress rate of the last decade, I’d expect it to take a century for research directions like these to actually get us to an understanding of minds sufficient to align them.

Maybe early breakthroughs chain into follow-up breakthroughs that shorten that time? Or maybe if you have fifty people trying that sort of thing, instead of 3–6, one of them ends up tugging on a thread that unravels the whole knot if they manage to succeed in time. It seems good to me that researchers are trying approaches like these, but the existence of a handful of people making such an attempt doesn’t seem to me to represent much of an update about humanity’s odds of survival.

High-level view

I again stress that all the people whose plans I am pessimistic about are people that I consider virtuous, and whose efforts I applaud. (And that my characterizations of people above are probably not endorsed by those people, and that I'm putting less effort into passing their ideological Turing Tests than would be virtuous of me, etc. etc.)

Nevertheless, my overall impression is that most of the new people coming into alignment research end up pursuing research that seems doomed to me, not just because they're unlikely to succeed at their stated research goals, but because their stated research goals have little overlap with what seem to me to be the tricky bits. Or, well, that's what happens at best; what happens at worst is they wind up doing capabilities work with a thin veneer of alignment research.

Perhaps unfairly, my subjective experience of people entering the alignment research field is that there are:

- a bunch of plans like Owen's (that seem to me to just completely miss the problem),
- and a bunch of people who study some local phenomenon of modern systems that seems to me to have little relationship to the difficult problems that I expect to arise once things start getting serious, while calling that "alignment" (thus watering down the term, and allowing them to convince themselves that alignment is actually easy because it's just as easy to train a language model to answer "morality" questions as it is to train it to explain jokes or whatever),
- and a few people who do capabilities work so that they can "stay near the action",
- and very few who are taking stabs at the hard problems.

An exception is interpretability work, which I endorse, and which I think is getting rightful efforts (though I will caveat that some grim part of me expects that somehow interpretability work will be used to boost capabilities long before it gets to the high level required to face down the tricky problems I expect in the late game). And there are definitely a handful of folk plugging away at research proposals that seem to me to have non-trivial inner product with the tricky problems.

In fact, when writing this list, I was slightly pleasantly surprised by how many of the research directions seem to me to have non-trivial inner product with the tricky problems.[6]

This isn't as much of a positive update as it might first seem, on account of how it looks to me like the total effort in the field is not distributed evenly across all the above proposals, and I still have a general sense that most researchers aren't really asking questions whose answers would really help us out. But it is something of a positive update nevertheless.

Returning to one of the better-by-my-lights proposals from above, Natural Abstractions: If this agenda succeeded and was correct in a key hypothesis, this would directly solve a big chunk of the problem.

I don't buy the key hypothesis (in the relevant way), and I don't expect that agenda to succeed.[7] But if I was saying that about a hundred pretty-uncorrelated agendas being pursued by two hundred people, I'd start to think that maybe the odds are in our favor.

My overall impression is still that when I actually look at the particular community we have, weighted by person-hours, the large majority of the field isn't trying to solve the problem(s) I expect to kill us. They're just wandering off in some other direction.

It could turn out that I’m wrong about one of these other directions. But "turns out the hard/deep problem I thought I could see, did not in fact exist" feels a lot less likely, on my own models, than "one of these 100 people, whose research would clearly solve the problem if it achieved its self-professed goals, might in fact be able to achieve their goals (despite me not sharing their research intuitions)".

So the status quo looks grim to me.

I in fact think it's nice to have *some* people saying "we can totally route around that problem", and then pursuing research paths that they think route around the problem!

But currently, we have only a few fractions of plans that look to me to be *trying* to solve the problem that I expect to *actually* kill us. Like a field of contingency plans with no work going into a Plan A; or like a field of pandemic preparedness that immediately turned its gaze away from the true disaster scenarios and focused the vast majority of its effort on ideas like “get people to eat healthier so that their immune systems will be better-prepared”. (Not a perfect analogy; sorry.)

Hence: I'm not highly-pessimistic about our prospects because I think this problem is extraordinarily hard. I think this problem is *normally* hard, and very little effort is being deployed toward solving it.

Like, you know how some people out there (who I'm reluctant to name for fear that reminding them of their old stances will contribute to fixing them in their old ways) are like, "Your mistake was attempting to put a goal into the AGI; what you actually need to do is keep your hands off it and raise it compassionately!"? And from our perspective, they're just walking blindly into the razor blades?

And then other people are like, "The problem is giving the AGI a bad goal, or letting bad people control it", and... well, that's probably still where some of you get off the train, but to the rest of us, these people *also* look like they're walking willfully into the razor blades?

Well, from my perspective, the people who are like, "Just keep training it on your objective while being somewhat clever about the training, maybe that empirically works", are also walking directly into the razor blades.

(And it doesn't help that a bunch of folks are like "Well, if you're right, then we'll be able to update later, when we observe that getting language models to answer ethical questions is mysteriously trickier than getting it to answer other sorts of questions", apparently impervious to my cries of "No, my model does not predict that, my model does not predict that we get all that much more advance evidence than we've got already". If the evidence we have isn't enough to get people focused on the central problems, then we seem to me to be in rather a lot of trouble.)

My current prophecy is not so much "death by problem too hard" as "death by problem not assailed".

Which is *absolutely* a challenge. I'd love to see more people attacking the things that seem to me like they're at the core.

1. ^

I ran a few of the dialogs past the relevant people, but that has empirically dragged out the amount of time it takes this post to publish, and I have a handful of other posts to publish afterwards, so I neglected to get feedback from most of the people mentioned. Sorry.

2. ^

   Much of Vanessa, Scott, etc.'s work does look to me like it is grappling with confusions related to the problem of aiming minds in theory, and if their research succeeds according to their own lights then I would expect to have a better understanding of how to aim minds in general, even ones that had undergone some sort of "sharp left turn".

   Which is not to say that I’m optimistic about *whether* any of these plans will succeed by their own lights. Regardless, they get points for taking a swing, and the thing I’m mostly advocating for is that more people take swings at this problem at all, not that we filter strongly on my optimism about specific angles of attack.

   I tried to solve the problem myself for a few years, and failed. Turns out I wasn't all that good at it.

   Maybe I'll be able to do better next time, and I poke at it every so often. (Even though in my mainline prediction, we won’t have the time to complete the sort of research paths that I can see and that I think have any chance of working.)

   MIRI funds or offers-to-fund most every researcher who I see as having this "their work would help with the generalization problem if they succeeded" property and as doing novel, nontrivial work, so it's no coincidence that I feel more positive about Vanessa, etc.'s work. But I'd like to see far more attempts to solve this problem than the field is currently marshaling.

3. ^

   Again, to be clear, it's nice to have some people trying to route around the hard problems wholesale. But I don't count such attempts as attacks on the problem itself. (I'm also not optimistic about any attempts I have yet seen to dodge the problem, but that's a digression from today's topic.)

4. ^

   I couldn't understand Stuart's views from what he's written publicly, so I ran this section by Stuart and Rebecca, who requested that I use actual quotes instead of my attempted paraphrasings. If I'd had more time, I'd like to have run all the dialogs by the researchers I mentioned in this post, and iterated until I could pass everyone's ideological Turing Test, as opposed to the current awkward set-up where the people that I thought I understood didn't get as much chance for feedback. But the time delay from editing this one section is evidence that this wouldn't be worth the time burnt. Instead, I hope the comments can correct any mischaracterizations on my part.

5. ^

   Note also that while having the AI ask for clarification in the face of ambiguity is nice and helpful, it is of course far from autonomous-AGI-grade.

6. ^

   I specifically see:

   - ~3 MIRI-supported research approaches that are trying to attack a chunk of the hard problem (with a caveat that I think the relevant chunks are too small and progress is too slow for this to increase humanity's odds of success by much).
   - ~1 other research approach that could maybe help address the core difficulty if it succeeds wildly more than I currently expect it to succeed (albeit no one is currently spending much time on this research approach): Natural Abstractions. Maybe 2, if you count sufficiently ambitious interpretability work.
   - ~2 research approaches that mostly don't help address the core difficulty (unless perhaps more ambitious versions of those proposals are developed, and the ambitious versions wildly succeed), but might provide small safety boosts on the mainline if other research addresses the core difficulty: Concept Extrapolation, and current interpretability work (with a caveat that sufficiently ambitious interpretability work would seem more promising to me than this).
   - 9+ approaches that appear to me to be either assuming away what look to me like the key problems, or hoping that we can do other things that allow us to avoid facing the problem: Truthful AI, ELK, AI Services, Evan's approach, the Richard/Rohin meta-approach, Vivek's approach, Critch's approach, superbabies, and the "maybe there is a pretty wide attractor basin around my own values" idea.
7. ^

   I rate "interpretability succeeds so wildly that we can understand and aim one of the first AGIs" as probably a bit more plausible than "natural abstractions are so natural that, by understanding them, we can practically find concepts-worth-optimizing-for in an AGI". Both seem very unlikely to me, though they meet my bar for “deserving of a serious effort by humanity” in case they work out.

The inordinately slow spread of good AGI conversations in ML

Spencer Greenberg wrote on Twitter:

> Recently @KerryLVaughan has been critiquing groups trying to build AGI, saying that by being aware of risks but still trying to make it, they're recklessly putting the world in danger. I'm interested to hear your thought/reactions to what Kerry says and the fact he's saying it.

Michael Page replied:

> I'm pro the conversation. That said, I think the premise -- that folks are aware of the risks -- is wrong.
>
> [...]
>
> Honestly, I think the case for the risks hasn't been that clearly laid out. The conversation among EA-types typically takes that as a starting point for their analysis. The burden for the we're-all-going-to-die-if-we-build-x argument is -- and I think correctly so -- quite high.

Oliver Habryka then replied:

> I find myself skeptical of this.
>
> [...]
>
> Like, my sense is that it's just really hard to convince someone that their job is net-negative. "It is difficult to get a man to understand something when his salary depends on his not understanding it" And this barrier is very hard to overcome with just better argumentation.

My reply:

I disagree with "the case for the risks hasn't been that clearly laid out". I think there's a giant, almost overwhelming pile of intro resources at this point, any one of which is more than sufficient, written in all manner of style, for all manner of audience.[1]

(I do think it's possible to create a much better intro resource than any that exist today, but 'we can do much better' is compatible with 'it's shocking that the existing material hasn't already finished the job'.)

I also disagree with "The burden for the we're-all-going-to-die-if-we-build-x argument is -- and I think correctly so -- quite high."

If you're building a machine, you should have an at least somewhat *lower* burden of proof for more serious risks. It's your responsibility to check your own work to some degree, and not impose lots of micromorts on everyone else through negligence.[2]

But I don't think the latter point matters much, since the 'AGI is dangerous' argument easily meets higher burdens of proof as well.

I do think a lot of people haven't heard the argument in any detail, and the main focus should be on trying to signal-boost the arguments and facilitate conversations, rather than assuming that everyone has heard the basics.

A lot of the field is very smart people who are stuck in circa-1995 levels of discourse about AGI.

I think 'my salary depends on not understanding it' is only a small part of the story. ML people could in principle talk way more about AGI, and understand the problem way better, without coming anywhere close to quitting their job. The level of discourse is by and large *too low* for 'I might have to leave my job' to be the very next obstacle on the path.

Also, many ML people have other awesome job options, have goals in the field other than pure salary maximization, etc.

More of the story: Info about AGI propagates too slowly through the field, because when one ML person updates, they usually don't loudly share their update with all their peers. This is because:

1. AGI sounds weird, and they don't want to sound like a weird outsider.

2. Their *peers* and the *community as a whole* might perceive this information as an attack on the field, an attempt to lower its status, etc.

3. Tech forecasting, differential technological development, long-term steering, exploratory engineering, 'not doing certain research because of its long-term social impact', prosocial research closure, etc. are very novel and foreign to most scientists.

EAs exert effort to try to dig up precedents like Asilomar partly *because* Asilomar is so unusual compared to the norms and practices of the vast majority of science. Scientists generally don't think in these terms at all, especially in *advance* of any major disasters their field causes.

And the scientists who do find any of this intuitive often feel vaguely nervous, alone, and adrift when they talk about it. On a gut level, they see that they have no institutional home and no super-widely-shared 'this is a virtuous and respectable way to do science' narrative.

Normal science is not Bayesian, is not agentic, is not 'a place where you're supposed to do arbitrary things just because you heard an argument that makes sense'. Normal science is a specific collection of scripts, customs, and established protocols.

In trying to move the field toward 'doing the thing that just makes sense', even though it's about a weird topic (AGI), and even though the prescribed response is also weird (closure, differential tech development, etc.), and even though the arguments in support are weird (where's the experimental data??), we're inherently fighting our way upstream, against the current.

Success is possible, but way, way more dakka is needed, and IMO it's easy to understand why we haven't succeeded more.

This is also part of why I've increasingly updated toward a strategy of "let's all be way too blunt and candid about our AGI-related thoughts".

The core problem we face isn't 'people informedly disagree', 'there's a values conflict', 'we haven't written up the arguments', 'nobody has seen the arguments', or even 'self-deception' or 'self-serving bias'.

The core problem we face is 'not enough information is transmitting fast enough, because people feel nervous about whether their private thoughts are in the Overton window'.

We need to throw a brick through the Overton window. Both by adopting a very general policy of candidly stating what's in our head, and by propagating the arguments and info a lot further than we have in the past. If you want to normalize weird stuff fast, you have to be weird.

Cf. *Inadequate Equilibria*:

> What broke the silence about artificial general intelligence (AGI) in 2014 wasn’t Stephen Hawking writing a careful, well-considered essay about how this was a real issue. The silence only broke when Elon Musk tweeted about Nick Bostrom’s *Superintelligence*, and then made an off-the-cuff remark about how AGI was “summoning the demon.”
>
> Why did that heave a rock through the Overton window, when Stephen Hawking couldn’t? Because Stephen Hawking *sounded like* he was trying hard to appear sober and serious, which signals that this is a subject you have to be careful not to gaffe about. And then Elon Musk was like, “*Whoa, look at that apocalypse over there!!*” After which there was the equivalent of journalists trying to pile on, shouting, “A gaffe! A gaffe! A… gaffe?” and finding out that, in light of recent news stories about AI and in light of Elon Musk’s good reputation, people weren’t backing them up on that gaffe thing.

> Similarly, to heave a rock through the Overton window on the War on Drugs, what you need is not state propositions (although those do help) or articles in *The Economist*. What you need is for some “serious” politician to say, “This is dumb,” and for the journalists to pile on shouting, “A gaffe! A gaffe… a gaffe?” But it’s a grave personal risk for a politician to test whether the public atmosphere has changed enough, and even if it worked, they’d capture very little of the human benefit for themselves.

Simone Sturniolo commented on "AGI sounds weird, and they don't want to sound like a weird outsider.":

> I think this is really the main thing. It sounds too sci-fi a worry. The "sensible, rational" viewpoint is that AI will never be that smart because haha, they get funny word wrong (never mind that they've grown to a point that would have looked like sorcery 30 years ago).

To which I reply: That's an example of a more-normal view that exists in society-at-large, but it's also a view that makes AI research sound lame. (In addition to being harder to say with a straight face if you've been working in ML for long at all.)

There's an important tension in ML between "play up AI so my work sounds important and impactful (and because it's in fact true)", and "downplay AI in order to sound serious and respectable".

This is a genuine tension, with no way out. There legit isn't any way to speak accurately and concretely about the future of AI without sounding like a sci-fi weirdo. So the field ends up tangled in ever-deeper knots motivatedly searching for some third option that doesn't exist.

Currently popular strategies include:

1. Quietism and directing your attention elsewhere.

2. Derailing all conversations about the future of AI to talk about semantics ("'AGI' is a wrong label").

3. Only talking about AI's long-term impact in extremely vague terms, and motivatedly focusing on normal goals like "cure cancer" since that's a normal-sounding thing doctors are already trying to do.

(Avoid any weird specifics about how you might go about curing cancer, and avoid weird specifics about the social effects of automating medicine, curing all disease, etc. Concreteness is the enemy.)

4. Say that AI's huge impacts will happen someday, in the indefinite future. But that's a "someday" problem, not a "soon" problem.

(Don't, of course, give specific years or talk about probability distributions over future tech developments, future milestones you expect to see 30 years before AGI, cruxes, etc. That's a weird thing for a scientist to do.)

5. Say that AI's impacts will happen gradually, over many years. Sure, they'll ratchet up to being a big thing, but it's not like any crazy developments will happen *overnight*; this isn't science fiction, after all.

(Somehow "does this happen in sci-fi?" feels to people like a relevant source of info about the future.)

When Paul Christiano talks about soft takeoff, he has in mind a scenario like 'we'll have some years of slow ratcheting to do some preparation, but things will accelerate faster and faster and be extremely crazy and fast in the endgame'.

But what people outside EA usually have in mind by soft takeoff is:

I think the Paul scenario is one where things start going crazy in the next few decades, and go more and more crazy, and are apocalyptically crazy in thirty years or so?

But what many ML people seemingly want to believe (or want to talk as though they believe) is a *Jetsons* world.

A world where we gradually ratchet up to "human-level AI" over the next 50–250 years, and then we spend another 50–250 years slowly ratcheting up to crazy superhuman systems.

The clearest place I've seen this perspective explicitly argued for is in Rodney Brooks' writing. But the much more common position isn't to explicitly argue for this view, or even to explicitly state it. It jut sort of lurks in the background, like an obvious sane-and-moderate Default. Even though it *immediately and obviously falls apart as a scenario* as soon as you start poking at it and actually discussing the details.

6. Just say it's not your job to think or talk about the future. You're a scientist! Scientists don't think about the future. They just do their research.

7. More strongly, you can say that it's irresponsible speculation to even broach the subject! What a silly thing to discuss!

Note that the argument here usually isn't "AGI is clearly at least 100 years away for reasons X, Y, and Z; therefore it's irresponsible speculation to discuss this until we're, like, 80 years into the future." Rather, *even giving arguments for why AGI is 100+ years away* is assumed at the outset to be irresponsible speculation. There isn't a cost-benefit analysis being given here for why this is low-importance; there's just a miasma of unrespectability.

1. ^

   Some of my favorite informal ones to link to: Russell 2014; Urban 2015 (w/ Muehlhauser 2015); Yudkowsky 2016a; Yudkowsky 2017; Piper 2018; Soares 2022

   Some of my favorite less-informal ones: Bostrom 2014a; Bostrom 2014b; Yudkowsky 2016b; Soares 2017; Hubinger et al. 2019; Ngo 2020; Cotra 2021; Yudkowsky 2022; Arbital's listing

   Other good ones include: Omohundro 2008; Yudkowsky 2008; Yudkowsky 2011; Muehlhauser 2013; Yudkowsky 2013; Armstrong 2014; Dewey 2014; Krakovna 2015; Open Philanthropy 2015; Russell 2015; Soares 2015a; Soares 2015b; Steinhardt 2015; Alexander 2016; Amodei et al. 2016; Open Philanthropy 2016; Taylor et. al 2016; Taylor 2017; Wiblin 2017; Yudkowsky 2017; Garrabrant and Demski 2018; Harris and Yudkowsky 2018; Christiano 2019a; Christiano 2019b; Piper 2019; Russell 2019; Shlegeris 2020; Carlsmith 2021; Dewey 2021; Miles 2021; Turner 2021; Steinhardt 2022

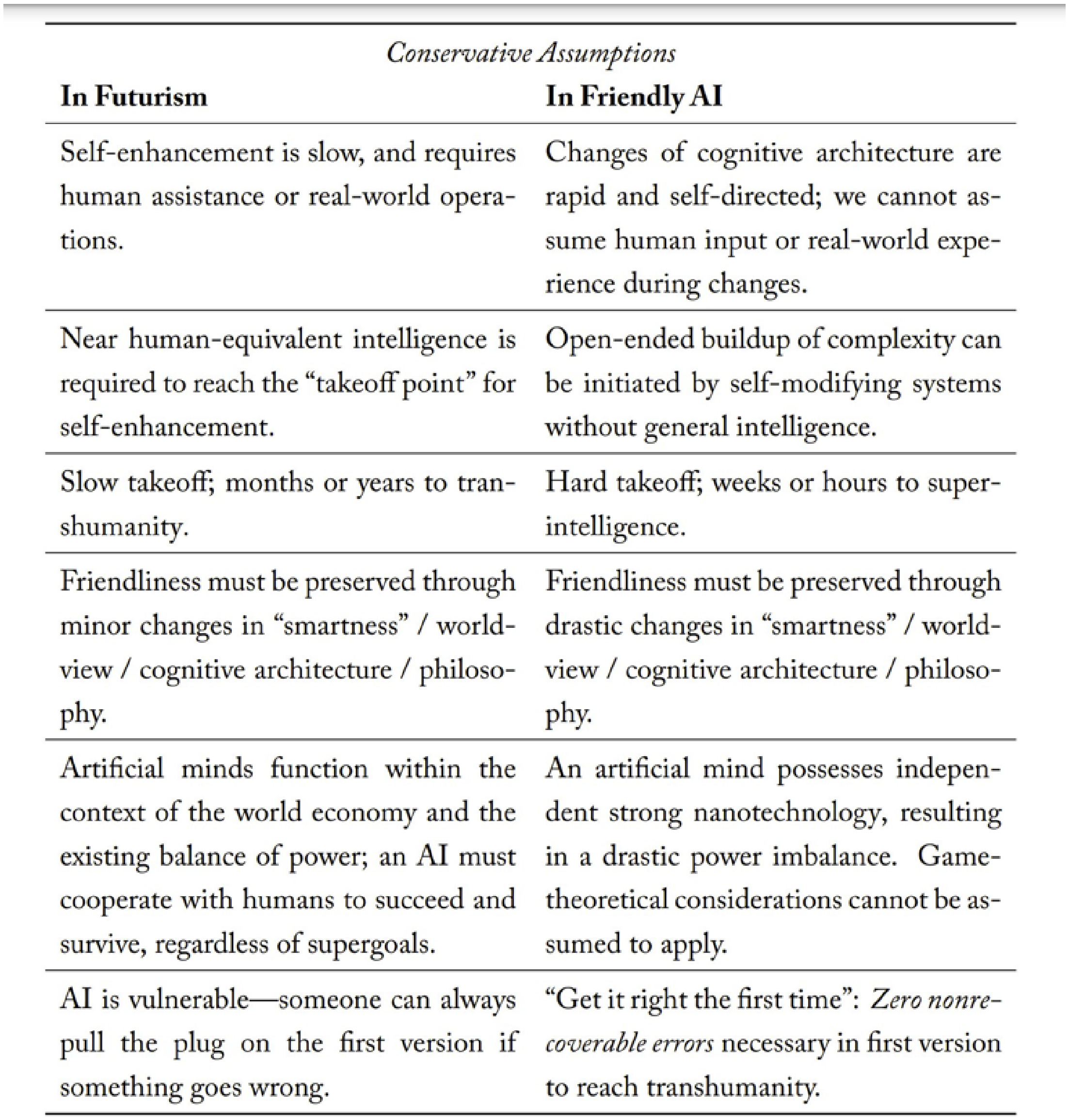
2. ^

   Or, I should say, a lower "burden of inquiry".

   You should (at least somewhat more readily) take the claim seriously and investigate it in this case. But you shouldn't *require less evidence* to *believe* anything — that would just be biasing yourself, unless you're already biased and are trying to debias yourself. (In which case this strikes me as a bad debiasing tool.)

   See also the idea of "conservative futurism" versus "conservative engineering" in *Creating Friendly AI 1.0*:

   > The conservative assumption according to futurism is not necessarily the “conservative” assumption in Friendly AI. Often, the two are diametric opposites. When building a toll bridge, the conservative *revenue* assumption is that half as many people will drive through as expected. The conservative *engineering* assumption is that ten times as many people as expected will drive over, and that most of them will be driving fifteen-ton trucks.
   >
   > Given a choice between discussing a human-dependent traffic-control AI and discussing an AI with independent strong nanotechnology, we should be biased towards assuming the more powerful and independent AI. An AI that remains Friendly when armed with strong nanotechnology is likely to be Friendly if placed in charge of traffic control, but perhaps not the other way around. (A minivan can drive over a bridge designed for armor-plated tanks, but not vice-versa.



Conservative Assumptions

In Futurism	In Friendly AI
Self-enhancement is slow, and requires human assistance or real-world operations.	Changes of cognitive architecture are rapid and self-directed; we cannot assume human input or real-world experience during changes.
Near human-equivalent intelligence is required to reach the "takeoff point" for self-enhancement.	Open-ended buildup of complexity can be initiated by self-modifying systems without general intelligence.
Slow takeoff; months or years to transhumanity.	Hard takeoff; weeks or hours to superintelligence.
Friendliness must be preserved through minor changes in "smartness" / worldview / cognitive architecture / philosophy.	Friendliness must be preserved through drastic changes in "smartness" / worldview / cognitive architecture / philosophy.
Artificial minds function within the context of the world economy and the existing balance of power; an AI must cooperate with humans to succeed and survive, regardless of supergoals.	An artificial mind possesses independent strong nanotechnology, resulting in a drastic power imbalance. Game-theoretical considerations cannot be assumed to apply.
AI is vulnerable—someone can always pull the plug on the first version if something goes wrong.	"Get it right the first time": *Zero nonrecoverable errors* necessary in first version to reach transhumanity.

The core argument for hard takeoff, 'AI can achieve strong nanotech', and "get it right the first time" is that they're *true*, not that they're "conservative". But it's of course also true that a sane world that thought hard takeoff were "merely" 20% likely, would not immediately give up and write off human survival in those worlds. Your plan doesn't need to survive one-in-a-million possibilities, but it should survive one-in-five ones!

A note about differential technological development

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Quick note: I occasionally run into arguments of the form "my research advances capabilities, but it advances alignment more than it advances capabilities, so it's good on net". I do not buy this argument, and think that in most such cases, this sort of research does more harm than good. (Cf. differential technological development.)

For a simplified version of my model as to why:

- Suppose that aligning an AGI requires 1000 person-years of research.
  - 900 of these person-years can be done in parallelizable 5-year chunks (e.g., by 180 people over 5 years — or, more realistically, by 1800 people over 10 years, with 10% of the people doing the job correctly half the time).
  - The remaining 100 of these person-years factor into four chunks that take 25 serial years apiece (so that you can't get any of those four parts done in less than 25 years).

In this toy model, a critical resource is *serial time*: if AGI is only 26 years off, then shortening overall timelines by 2 years is a death sentence, *even if you're getting all 900 years of the "parallelizable" research done in exchange.*

My real model of the research landscape is more complex than this toy picture, but I do in fact expect that serial time is a key resource when it comes to AGI alignment.

The most blatant case of alignment work that seems **parallelizable** to me is that of "AI psychologizing": we can imagine having enough success building comprehensible minds, and enough success with transparency tools, that with a sufficiently large army of people studying the alien mind, we can develop a pretty good understanding of what and how it's thinking. (I currently doubt we'll get there in practice, but if we did, I could imagine most of the human-years spent on alignment-work being sunk into understanding the first artificial mind we get.)

The most blatant case of alignment work that seems **serial** to me is work that requires having a theoretical understanding of minds/optimization/whatever, or work that requires having just the right concepts for thinking about minds. Relative to our current state of knowledge, it seems to me that a lot of serial work is plausibly needed in order for us to understand how to safely and reliably aim AGI systems at a goal/task of our choosing.

A bunch of modern alignment work seems to me to sit in some middle-ground. As a rule of thumb, alignment work that is closer to behavioral observations of modern systems is more parallelizable (because you can have lots of people making those observations in parallel), and alignment work that requires having a good conceptual or theoretical framework is more serial (because, in the worst case, you might need a whole new generation of researchers raised with a half-baked version of the technical framework, in order to get people who both have enough technical clarity to grapple with the remaining confusions, and enough youth to invent a whole new way of seeing the problem—a pattern which seems common to me in my read of the development of things like analysis, meta-mathematics, quantum physics, etc.).

As an egregious and fictitious (but "based on a true story") example of the arguments I disagree with, consider the following dialog:

Uncharacteristically conscientious capabilities researcher: Alignment is made significantly trickier by the fact that we do not have an artificial mind in front of us to study. By doing capabilities research now (and being personally willing to pause when we get to the brink), I am making it more possible to do alignment research.

Me: Once humanity gets to the brink, I doubt we have much time left. (For a host of reasons, including: simultaneous discovery; the way the field seems to be on a trajectory to publicly share most of the critical AGI insights, once it has them, before wisening up and instituting closure policies after it's too late; Earth's generally terrible track-record in cybersecurity; and a sense that excited people will convince themselves it's fine to plow ahead directly over the cliff-edge.)

Uncharacteristically conscientious capabilities researcher: Well, we might not have many *sidereal* years left after we get to the brink, but we'll have many, *many* more *researcher* years left. The top minds of the day will predictably be much more interested in alignment work when there's an actual misaligned artificial mind in front of them to study. And people will take these problems much more seriously once they're near-term. And the monetary incentives for solving alignment will be much more visibly present. And so on and so forth.

Me: Setting aside how I believe that the world is derpier than that: even if you were right, I still think we'd be screwed in that scenario. In particular, that scenario seems to me to assume that there is not much serial research labor needed to do alignment research.

Like, I think it's quite hard to get something akin to Einstein's theory of general relativity, or Grothendieck's simplification of algebraic geometry, without having some researcher retreat to a mountain lair for a handful of years to build/refactor/distill/reimagine a bunch of the relevant concepts.

And looking at various parts of the history of math and science, it looks to me like technical fields often move forwards by building up around subtly-bad framings and concepts, so that a next generation can be raised with enough technical machinery to grasp the problem and enough youth to find a whole new angle of attack, at which point new and better framings and concepts are invented to replace the old. "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die and a new generation grows up that is familiar with it" (Max Planck) and all that.

If you need the field to iterate in that sort of way three times before you can see clearly enough to solve alignment, you're going to be hard-pressed to do that in five years no matter how big and important your field seems once you get to the brink.

(Even the 25 years in the toy model above feels pretty fast, to me, for that kind of iteration, and signifies my great optimism in what humanity is capable of doing in a rush when the whole universe is on the line.)

It looks to me like alignment requires both a bunch of parallelizable labor and a bunch of serial labor. I expect us to have very little serial time (a handful of years if we're lucky) after we have fledgling AGI.

When I've heard the “two units of alignment progress for one unit of capabilities progress” argument, my impression is that it's been made by people who are burning *serial* time in order to get a bit more of the *parallelizable* alignment labor done.

But the parallelizable alignment labor is not the bottleneck. The serial alignment labor is the bottleneck, and it looks to me like burning time to complete *that* is nowhere near worth the benefits in practice.

Some nuance I'll add:

I feel relatively confident that a large percentage of people who do capabilities work at OpenAI, FAIR, DeepMind, Anthropic, etc. with justifications like "well, I'm helping with alignment some too" or "well, alignment will be easier when we get to the brink" (more often EA-adjacent than centrally "EA", I think) are currently producing costs that outweigh the benefits.

Some relatively niche and theoretical agent-foundations-ish research directions might yield capabilities advances too, and I feel much more positive about those cases. I’m guessing it won’t *work*, but it’s the kind of research that seems positive-EV to me and that I’d like to see a larger network of researchers tackling, provided that they avoid publishing large advances that are especially likely to shorten AGI timelines.

The main reasons I feel more positive about the agent-foundations-ish cases I know about are:

- The alignment progress in these cases appears to me to be much more serial, compared to the vast majority of alignment work the field outputs today.
- I’m more optimistic about the *total amount* of alignment progress we’d see in the worlds where agent-foundations-ish research so wildly exceeded my expectations that it ended up boosting capabilities. Better understanding optimization in this way really would seem to me to take a significant bite out of the capabilities generalization problem, unlike most alignment work I’m aware of.
- The kind of people working on agent-foundations-y work aren’t publishing new ML results that break SotA. Thus I consider it more likely that they’d avoid publicly breaking SotA on a bunch of AGI-relevant benchmarks given the opportunity, and more likely that they’d only direct their attention to this kind of intervention if it seemed helpful for humanity’s future prospects.[1]
- Relatedly, the energy and attention of ML is elsewhere, so if they do achieve a surprising AGI-relevant breakthrough and accidentally leak bits about it publicly, I put less probability on safety-unconscious ML researchers rushing to incorporate it.

I’m giving this example not to say “everyone should go do agent-foundations-y work exclusively now!”. I think it’s a neglected set of research directions that deserves far more effort, but I’m far too pessimistic about it to want humanity to put all its eggs in that basket.

Rather, my hope is that this example clarifies that I’m not saying “doing alignment research is bad” or even “all alignment research that poses a risk of advancing capabilities is bad”. I think that in a large majority of scenarios where humanity’s long-term future goes well, it mainly goes well because we made major alignment progress over the coming years and decades.[2] I don’t want this post to be taken as an argument against what I see as humanity’s biggest hope: figuring out AGI alignment.

1. ^

   On the other hand, weirder research is more likely to shorten timelines a *lot*, if it shortens them at all. More mainstream research progress is less likely to have a large counterfactual impact, because it’s more likely that someone else has the same idea a few months or years later.

   “Low probability of shortening timelines a lot” and “higher probability of shortening timelines a smaller amount” both matter here, so I advocate that both niche and mainstream researchers be cautious and deliberate about publishing potentially timelines-shortening work.

2. ^

   "Decades" would require timelines to be longer than my median. But when I condition on success, I do expect we have more time.

Brainstorm of things that could force an AI team to burn their lead

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

> **Comments:** The following is a list (very lightly edited with help from Rob Bensinger) I wrote in July 2017, at Nick Beckstead's request, as part of a conversation we were having at the time. From my current vantage point, it strikes me as narrow and obviously generated by one person, listing the first things that came to mind on a particular day.
>
> I worry that it's easy to read the list below as saying that this narrow slice, all clustered in one portion of the neighborhood, is a very big slice of the space of possible ways an AGI group may have to burn down its lead.
>
> This is one of my models for how people wind up with really weird pictures of MIRI beliefs. I generate three examples that are clustered together because I'm bad at generating varied examples on the fly, while hoping that people can generalize to see the broader space these are sampled from; then people think I've got a fetish for the particular corner of the space spanned by the first few ideas that popped into my head. E.g., they infer that I must have a bunch of other weird beliefs that force reality into that particular corner.
>
> I also worry that the list below doesn't come with a sufficiently loud disclaimer about how the real issue is earlier and more embarrassing. The real difficulty isn't that you make an AI and find that it's mostly easy to align except that it happens to befall issues b, d, and g. The thing to expect is more like: you just have this big pile of tensors, and the interpretability tools you've managed to scrounge together give you flashes of visualizations of its shallow thoughts, and the thoughts say "yep, I'm trying to kill all humans", and you are just utterly helpless to do anything about, because you don't have the sort of mastery of its cognition that you'd need to reach in and fix that and you wouldn't know how to fix it if you did. And you have nothing to train against, except the tool that gives you flashes of visualizations (which would just train fairly directly against interpretability, until it was thinking about how to kill all humans somewhere that you couldn't see).
>
> The brainstormed list below is an exercise in how, if you zoom in on any part of the problem, reality is just allowed to say "lol nope" to you from many different angles simultaneously. It's intended to convey some of the difference (that every computer programmer knows) between "I can just code X" and "wow, there is a lot of subtlety to getting X right"; the difference between the optimistic hope in-advance that everything is going to go smoothly, and the excessively detailed tarpit of reality. This is not to be confused with thinking that these hurdles are a particularly representative sample, much less an attempt to be exhaustive.

Context

The imaginary group DeepAI pushed to get an AGI system as fast as reasonably possible. They now more or less understand how to build something that is very good at generalized learning and cross-domain reasoning and what-not. They rightfully believe that, if they had a reckless desire to increase the capabilities of the system as fast as possible without regard for the consequences, they would be able to have it recursively self-improving within a year. However, their existing system is not yet a superintelligence, and does not yet have the resources to be dangerous in its own right.

For the sake of concreteness, we will imagine that the system came largely from an extension of modern AI techniques: a large amount of end-to-end training, heavy use of neural networks, heavy use of reinforcement learning, and so on.

The question is, what sorts of things might they discover about the system that force them to stop and redesign (and/or recode, and/or retrain) large parts of the system?

Brainstorm list

(Note: Bullet points are highly disjunctive. Also, I'm leaning on the side of telling evocative stories so as to increase the chance of getting the point across; obviously, each specific detail is burdensome, and in each case I'm trying to wave in the direction of a more general class of possible failures. Also, to state the obvious, this list does not feel complete to me, and I find some of these points to be more plausible than others.)

- (a) They want to put in alarms that warn them when the system is thinking a class of thought that they don't want thought, but…
    - the system's analog of "thought processes" are not amenable to programmatic classification, because…
        - the "thoughts" are so opaque that the programmers cannot figure them out for quite some time.
        - the representation / data structure is convoluted, and simple classification systems can't figure it out (in the same way that a modern narrow AI system can understand sentiment but not content of a science paper).
        - the "thoughts" are not centralized; they arise out of interactions between many scattered parts of the system and an extensive redesign is required to make it possible to collate them and expose them to automated tools.
        - the system has internal control of its own "thought language", and it changes rapidly enough that narrower automated tools can't keep up; there is no easy way to slow down the shift to its internal thought-speak without crippling it.
    - the system simply wasn't designed for monitoring of this form, and…

        - the code must be heavily refactored in order to even allow the relevant data about the system’s thoughts to be collected in a useful fashion.
        - the code must be heavily refactored in order to allow live monitors and checks to be attached in a way that do not cause an intolerable slowdown.

- (b) They want to blacklist some domain of reasoning (either for alignment reasons or because the system is getting confused by irrelevant reasoning that they want to cut out); or they want to whitelist a set of reasoning domains; and the system simply was not designed to allow this.
    - Simple attempts to blacklist a domain result in nearest-unblocked-strategy problems. Solving the problem at the root requires re-architecting the system and a significant amount of retraining.
    - More sophisticated attempts to blacklist a single domain cripple the entire system. For example, it isn’t supposed to think about ways to deceive humans, and this destroys its ability to ask clarifying questions of the programmers.
    - Or, worse, the system is such a mess of spaghetti that when you try to prevent it from thinking too hard about geopolitics, for indecipherable reasons, it stops being able to think at all. (Later it was discovered that some crucial part of the system was figuring out how to manage some crucial internal resource by having some other part of the system think about hypothetical "geopolitics" questions, because what did you expect, your AGI’s internals are a mess.)

- (c) The operators realize that the system’s internal objectives are not lining up with their own objectives. This is very difficult for them to fix, because…
    - the system achieved its high performance by being walked through a large number of objectives in heavily reward-landscaped environments (generated by large amounts of data). The system now has the world-models and the capabilities to pursue ambitious real-world objectives, but the only interface that the programmers have by which to point at an objective is via reward-landscaped objective functions generated by mountains of data. This is no longer sufficient, because…
        - the tasks at hand are not amenable to the generation of large amounts of data (e.g., we can’t generate a nicely landscaped reward function between here and “nanofabricator”, and we don’t have many examples of not-quite-nanofabricators to provide). The show is stopped.
        - the system has no interface through which the programmers can sift through the concepts in its world-model and pick out (or create, in something sufficiently close to the system’s native tongue for this to be fine) the concept corresponding to “nanofabricator”. Exposing that interface requires significant refactoring and some redesign.
        - the system’s concept language is opaque, and the programmers keep picking out something that’s not quite the nanofabricator concept, and the system keeps going down wrong paths. Developing translation tools for the system’s internal concept language requires significant time and effort.

        - the internal concept language is constantly in flux; causing it to stay fixed long enough for the programmers to pick out a goal requires significant effort.
        - the programmers have no mechanism for tying a selected concept into the system’s main optimization procedures. The system is very good at optimizing in a way that causes rewards (or whatever the analog of reward it was trained on) to be high, but given a pointer to a certain concept in the system, it is not apparent how to design a reward landscape that makes the system optimize for a chosen concept.

            And this is exacerbated by the fact that the system has no centralized optimization procedure; it instead has a large collection of internal processes that interact in a way that causes the predicted rewards to be high, but it is very difficult to identify and understand all those internal processes sufficiently well to get them all pointed at something other than optimizing in favor of the reward channel.

            Their attempts keep failing because, e.g., subsystem X had a heuristic to put its outputs in location Y, which is where subsystem Z would have been looking for them if subsystem Z had been optimizing the reward channel, but optimization of some other arbitrary concept causes Z’s “look in location Y” heuristic to become invalidated for one reason or another, and that connection stops occurring. And so on and so forth; aligning all the internal subprocesses to pursue something other than the reward channel proves highly difficult.
    - the system is having a particularly hard time learning the boundaries of the human concept: its empirically-motivated internal language does not beget short descriptions of value-laden objectives. Significant retraining is required to develop a language in which it can even develop the concept of the goal.

- (d) In order to get the system to zero in on the operators’ goals, they decide to have the system ask the humans various questions at certain key junctures. This proves more difficult than expected, because…
    - the system wasn’t designed to allow this, and it’s pretty hard to add all the right hooks (for similar reasons to why it might be difficult to add alarms).
    - the system vacillates between asking far too many and far too few questions, and a lot of thought and some redesign/retraining is necessary in order to get the question-asking system to the point where the programmers think it might actually provide the desired safety coverage.
    - the system does not yet have an understanding of human psychology sufficient for it to be able to ask the right questions in value-laden domains, and significant time is wasted trying to make this work when it can’t.
    - relatedly, the system is not yet smart enough to generalize over the human answers in a reasonable fashion, causing it to gain far less from the answers than humans think it should, and solving this would require ramping up the system’s capabilities to an unsafe level.
    - the system has no mechanism for translating its more complex / complicated / subtle questions into questions that humans can understand

and provide reasonable feedback on. Fixing this requires many months of effort, because…

    - understanding the questions well enough to even figure out how to translate them is hard.
    - building the translation tool is hard.
    - the system is bad at describing the likely consequences of its actions in human-comprehensible terms. Fixing this is hard for, e.g., reasons discussed under (c).

- (e) The early system is highly goal-directed through and through, and the developers want to switch to something more like “approval direction all the way down”. This requires a large and time-intensive refactor (if it's even reasonably-possible at all).

- (f) Or, conversely, the system starts out a mess, and the developers want to switch to a “goal directed all the way down” system, where every single computation in the system is happening for a known purpose (and some other system is monitoring and making sure that every subprocess is pursuing a particular narrow purpose). Making this possible requires a time-intensive refactor.

- (g) The programmers want to remove all “argmaxing” (cases of unlimited optimization inside the system, such as “just optimize the memory efficiency as hard as possible”). They find this very difficult for reasons discussed above (the sources of argmaxing behavior are difficult to identify; limiting an argmax in one part of the system breaks some other far-flung part of the system for difficult-to-decipher reasons; etc. etc. etc.).

- (h) The programmers want to track how much resource the system is putting towards various different internal subgoals, but this is difficult for reasons discussed above, etc.

- (i) The programmers want to add any number of other safety features (limited impact, tripwires, etc.) and find this difficult for reasons listed above, etc.

- (j) The internal dynamics of the system are revealed to implement any one of a bajillion false dichotomies, such as “the system can either develop reasonable beliefs about X, or pursue goal Y, but the more we improve its beliefs about X the worse it gets at pursuing Y, and vice versa.” (There are certainly human cases in human psychology where better knowledge of fact X makes the human less able to pursue goal Y, and this seems largely silly.)

- (k) Generalizing over a number of points that appeared above, the programmers realize that they need to make the system broadly more…
    - transparent. Its concepts/thought patterns are opaque black boxes. They've burned time understanding specific types of thought patterns in many specific instances, and now they have some experience with the system, and want to refactor/redesign/retrain such that it's more transparent across the board. This requires a number of months.
    - debuggable. Its internals are interdependent spaghetti, where (e.g.) manually modifying a thought-suggesting system to add basic alarm systems violates assumptions that some other far-flung part of the system was depending on; this is a pain in the ass to debug. After a number of these issues arise, the programmers decide that they cannot safely proceed until they…
        - cleanly separate various submodules by hand, and to hell with end-to-end training. This takes many months of effort.
        - retrain the system end-to-end in a way that causes its internals to be more modular and separable. This takes many months of effort.

- (l) Problems crop up when they try to increase the capabilities of the system. In particular, the system…
    - finds new clever ways to wirehead.
    - starts finding "epistemic feedback loops" such as the Santa clause sentence ("If this sentence is true, then Santa Claus exists") that, given it's internally hacky (and not completely sound) reasoning style, allow it to come to any conclusion if it thinks the right thoughts in the right pattern.
    - is revealed to have undesirable basic drives (such as a basic drive for efficient usage of memory chips), in a fashion similar to how humans have a basic drive for hunger, in a manner that affects its real-world policy suggestions in a sizable manner. While the programmers have alarms that notice this and go off, it is very deep-rooted and particularly difficult to remove or ameliorate without destroying the internal balance that causes the system to work at all.
        - The system develops a reflective instability. For example, the system previously managed its internal resources by spawning internal goals for things like scheduling and prioritization, and as the system scales and gets new, higher-level concepts, it regularly spawns internal goals for large-scale self-modifications which it would not be safe to allow. However, preventing these proves quite difficult, because…
            - detecting them is tough.
            - manually messing with the internal goal system breaks everything.
            - nearest-unblocked-strategy problems.
        - It realizes that it has strong incentives to outsource its compute into the external environment. Removing this is difficult for reasons discussed above.
        - Subprocesses that were in delicate balance at capability level X fall out of balance as capabilities are increased, and a single module begins to dominate the entire system.
            - For example, maybe the system uses some sort of internal market economy for allocating credit, and as the resources ramp up, certain cliques start to get a massive concentration of "wealth" that causes the whole system to gum up, and this is

difficult to understand / debug / fix because the whole thing was so delicate in the first place.

- (m) The system is revealed to have any one of a bajillion cognitive biases often found in humans, and it’s very difficult to track down why or to fix it, but the cognitive bias is sufficient to make the system undeployable.
    - Example: it commits a variant of the sour grapes fallacy where whenever it realizes that a goal is difficult it updates both its model of the world and its preferences about how good it would be to achieve that goal; this is very difficult to patch because the parts of the system that apply updates based on observation were end-to-end trained, and do not factor nicely along “probability vs utility” lines.

- (n) The system can be used to address various issues of this form, but only by giving it the ability to execute unrestricted self-modification. The extent, rapidity, or opacity of the self-modifications are such that humans cannot feasibly review them. The design of the system does not allow the programmers to easily restrict the domain of these self-modifications such that they can be confident that they will be safe. Redesigning the system such that it can fix various issues in itself without giving it the ability to undergo full recursive self-improvement requires significant redesign and retraining.

- (o) As the team is working to get the system deployment-ready for some pivotal action, the system’s reasoning is revealed to be corrupted by flaws in some very base-level concepts. The system requires significant retraining time and some massaging on the code/design levels in order to change these concepts and propagate some giant updates; this takes a large chunk of time.

- (p) The system is very easy to fool, trick, blackmail, or confuse-into-revealing-all-its-secrets, or similar. The original plan that the operators were planning to pursue requires putting the system out in the environment where adversarial humans may attempt to take control of the system or otherwise shut it down. Hardening the system against this sort of attack requires many months of effort, including extensive redesign/retraining/recoding.

- (q) The strategy that the operators were aiming for requires cognitive actions that the programmers eventually realize is untenable in the allotted time window or otherwise unsafe, such as deep psychological modeling of humans. The team eventually decides to choose a new pivotal action to target, and this new strategy requires a fair bit of redesign, recoding, and/or retraining.

Asides

- My impression is that most catastrophic bugs in the space industry are not due to code crashes / failures; they are instead due to a normally-reliable module producing a wrong-but-syntactically-close-to-right valid-seeming output at an inopportune time. It seems very plausible to me that first-pass AGI systems will be in the category of things that work via dividing labor across a whole bunch of interoperating internal modules; insofar as errors can cascade when a normally-reliable module outputs a wrong-but-right-seeming output at the wrong time, I think we do in fact need to treat “getting the AGI’s internals right” as being in the same reference class as “get the space probe’s code right”.
- Note, as always, that detecting the problem is only half the battle – in all the cases above, I’m not trying to point and say “people might forget to check this and end the world”; rather, I’m saying, “once this sort of error is detected, I expect that the team will need to burn a chunk of time to correct it”.
- Recall that this is a domain where playing whack-a-mole gets you killed: if you have very good problem-detectors, and you go around removing problem symptoms instead of solving the underlying root problem, then eventually your problem-detectors will stop going off, but this will not be because your AGI is safe to run. In software, removing the symptoms is usually way easier than fixing a problem at the root cause; I worry that fixing these sorts of problems at their root cause can require quite a bit of time.
- Recall that it’s far harder to add a feature to twitter than it is to add the same feature to a minimalistic twitter clone that you banged out in an afternoon. Similarly, solving an ML problem in a fledgling AGI in a way that integrates with the rest of the system without breaking anything delicate is likely way harder than solving an analogous ML problem in a simplified setting from a clean slate.

Finally, note that this is only intended as a brainstorm of things that might force a leading team to burn a large number of months; it is not intended to be an exhaustive list of reasons that alignment is hard. (That would include various other factors such as “what sorts of easy temptations will be available that the team has to avoid?” and “how hard is it to find a viable deployment strategy?” and so on.)

AGI ruin scenarios are likely (and disjunctive)

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Note: As usual, Rob Bensinger helped me with editing. I recently discussed this model with Alex Lintz, who might soon post his own take on it (edit: here).

Some people seem to be under the impression that I believe AGI ruin is a small and narrow target to hit. This is not so. My belief is that *most* of the outcome space is full of AGI ruin, and that *avoiding* it is what requires navigating a treacherous and narrow course.

So, to be clear, here is a very rough model of why I think AGI ruin is likely. (>90% likely in our lifetimes.)[1]

My real models are more subtle, take into account more factors, and are less articulate. But people keep coming to me saying "it sounds to me like you think humanity will somehow manage to walk a tightrope, traverse an obstacle course, and thread a needle in order to somehow hit the narrow target of catastrophe, and I don't understand how you're so confident about this". (Even after reading Eliezer's AGI Ruin post—which I predominantly agree with, and which has a very disjunctive character.)

Hopefully this sort of toy model will at least give you some vague flavor of where I'm coming from.

Simplified Nate-model

The short version of my model is this: from the current position on the game board, a lot of things need to go right, if we are to survive this.

In somewhat more detail, the following things need to go right:

- The world's overall state needs to be such that AI can be deployed to make things good. A non-exhaustive list of things that need to go well for this to happen follows:
    - The world needs to admit of an AGI deployment strategy (compatible with realistic alignable-capabilities levels for early systems) that prevents the world from being destroyed if executed.
    - At least one such strategy needs to be known and accepted by a leading organization.
    - Somehow, at least one leading organization needs to have enough time to nail down AGI, nail down alignable AGI, actually build+align their system, and deploy their system to help.
        - This very likely means that there needs to either be only one organization capable of building AGI for several years, or all the AGI-

capable organizations need to be very cautious and friendly and deliberately avoid exerting too much pressure upon each other.

  - It needs to be the case that no local or global governing powers flail around (either prior to AGI, or during AGI development) in ways that prevent a (private or public) group from saving the world with AGI.
- Technical alignment needs to be solved to the point where good people could deploy AI to make things good. A non-exhaustive list of things that need to go well for this to happen follows:
  - There need to be people who think of themselves as working on technical alignment, whose work is integrated with AGI development and is a central input into how AGI is developed and deployed.
  - They need to be able to perceive every single lethal problem far enough in advance that they have time to solve them.
  - They need to be working on the problems in a way that is productive.
  - The problems (and the general paradigm in which they're attacked) need to be such that people's work can stack, or such that they don't require much serial effort; or the research teams need a lot of time.
  - Significant amounts of this work have to be done without an actual AGI to study and learn from; or the world needs to be able to avoid deploying misaligned AGI long enough for the research to complete.
- The internal dynamics at the relevant organizations need to be such that the organizations deploy an AGI to make things good. A non-exhaustive list of things that need to go well for this to happen follows:
  - The teams that first gain access to AGI, need to care in the right ways about AGI alignment.
    - E.g., they can't be "just raise the AGI with kindness; any attempt to force our values on it will just make it hate us" style kooks, or any other variety of kook you care to name.
  - The internal bureaucracy needs to be able to distinguish alignment solutions from fake solutions, quite possibly over significant technical disagreement.
    - This ability very likely needs to hold up in the face of immense social and time pressure.
  - People inside the organization need to be able to detect dangerous warning signs.
  - Those people might need very large amounts of social capital inside the organization.
  - While developing AGI, the team needs to avoid splintering or schisming in ways that result in AGI tech proliferating to other organizations, new or old.
  - The team otherwise needs to avoid (deliberately or accidentally) leaking AGI tech to the rest of the world during the development process.
  - The team likewise needs to avoid leaking insights to the wider world *prior* to AGI, insofar as accumulating proprietary insights enables the group to have a larger technical lead, and insofar as a larger technical lead makes it possible for you to e.g. have three years to figure out alignment once you reach AGI, as opposed to six months.

(I could also add a list of possible disasters from misuse, conditional on us successfully navigating all of the above problems. But conditional on us clearing all of the above hurdles, I feel pretty optimistic about the relevant players' reasonableness, such that the remaining risks seem much more moderate and tractable to my eye. Thus I'll leave out misuse risk from my AGI-ruin model in this post; e.g., the ">90% likely in our lifetimes" probability is just talking about misalignment risk.)

One way that this list is a toy model is that it’s assuming we have an actual alignment problem to face, under some amount of time pressure. Alternatives include things like getting (fast, high-fidelity) whole-brain emulation before AGI (which comes with a bunch of its own risks, to be clear). The probability that we somehow dodge the alignment problem in such a way puts a floor on how low models like the above can drive the probabilities of success down (though I’m pessimistic enough about the known-to-me non-AGI strategies that my unconditional p(ruin) is nonetheless >90%).

Some of these bullets trade off against each other: sufficiently good technical solutions might obviate the need for good AGI-team dynamics or good global-scale coordination, and so on. So these factors aren't totally disjunctive. But this list hopefully gives you a flavor for how it looks to me like a lot of separate things need to go right, simultaneously, in order for us to survive, at this point. Saving the world requires threading the needle; destroying the world is the default.

Correlations and general competence

You may object: "But Nate, you've warned of the multiple-stage fallacy; surely here you're guilty of the dual fallacy? You can't say that doom is high because three things need to go right, and multiply together the lowish probabilities that all three go right individually, because these are probably correlated."

Yes, they are correlated. They're especially correlated through the fact that the world is derpy.

This is the world where the US federal government's response to COVID was to ban private COVID testing, confiscate PPE bought by states, and warn citizens not to use PPE. It's a world where most of the focus on technical AGI alignment comes from our own local community, takes up a tiny fraction of the field, and most of it doesn't seem to me to be even trying by their own lights to engage with what look to me like the lethal problems.

Some people like to tell themselves that surely we'll get an AI warning shot and that will wake people up; but this sounds to me like wishful thinking from the world where the world has a competent response to the pandemic warning shot we just got.

So yes, these points are correlated. The ability to solve one of these problems is evidence of ability to solve the others, and the good news is that no amount of listing out more problems can drive my probability lower than the probability that I'm simply wrong about humanity's (future) competence. Our survival probability is greater than the product of the probability of solving each individual challenge.

The bad news is that we seem pretty deep in the competence-hole. We are not one mere hard shake away from everyone snapping to our sane-and-obvious-feeling views. You shake the world, and it winds up in some even stranger state, not in your favorite state.

(In the wake of the 2012 US presidential elections, it looked to me like there was clearly pressure in the US electorate that would need to be relieved, and I was cautiously optimistic that maybe the pressure would force the left into some sort of atheistish torch-of-the-enlightenment party and the right into some sort of libertarian individual-rights party. I, uh, wasn't wrong about there being pressure in the US electorate, but, the 2016 US presidential elections were not exactly what I was hoping for. But I digress.)

Regardless, there's a more general sense that a lot of things need to go right, from here, for us to survive; hence all the doom. And, lest you wonder what sort of single correlated already-known-to-me variable could make my whole argument and confidence come crashing down around me, it's whether humanity's going to rapidly become much more competent about AGI than it appears to be about everything else.

(This seems to me to be what many people imagine will happen to the pieces of the AGI puzzle other than the piece they’re most familiar with, via some sort of generalized Gell-Mann amnesia: the tech folk know that the technical arena is in shambles, but imagine that policy has the ball, and vice versa on the policy side. But whatever.)

So that's where we get our remaining probability mass, as far as I can tell: there's some chance I'm wrong about humanity's overall competence (in the nearish future); there's some chance that this whole model is way off-base for some reason; and there's a teeny chance that we manage to walk this particular tightrope, traverse this particular obstacle course, and thread this particular needle.

And again, I stress that the above is a toy model, rather than a full rendering of all my beliefs on the issue. Though my real model does say that a bunch of things have to go right, if we are to succeed from here.

1. ^

   In stark contrast to the multiple people I’ve talked to recently who thought I was arguing that there's a small chance of ruin, but the *expected* harm is so large as to be worth worrying about. *No*.

Where I currently disagree with Ryan Greenblatt’s version of the ELK approach

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Context: This post is my attempt to make sense of Ryan Greenblatt's research agenda, as of April 2022. I understand Ryan to be heavily inspired by Paul Christiano, and Paul left some comments on early versions of these notes.

Two separate things I was hoping to do, that I would have liked to factor into two separate writings, were (1) translating the parts of the agenda that I understand into a format that is comprehensible to me, and (2) distilling out conditional statements we might all agree on (some of us by rejecting the assumptions, others by accepting the conclusions). However, I never got around to that, and this has languished in my drafts folder too long, so I'm lowering my standards and putting it out there.

The process that generated this document is that Ryan and I bickered for a while, then I wrote up what I understood and shared it with Ryan, and we repeated this process a few times. I've omitted various intermediate drafts, on the grounds that sharing a bunch of intermediate positions that nobody endorses is confusing (moreso than seeing more of the process is enlightening), and on the grounds that if I try to do something better then what happens instead is that the post languishes in the drafts folder for half a year.

(Thanks to Ryan, Paul, and a variety of others for the conversations.)

Nate's model towards the end of the conversation

Ryan’s plan, as Nate currently understands it:

- Assume AGI is going to be paradigmatic, in the sense of being found by something roughly like gradient descent tuning the parameters in some fixed architecture. (This is not intended to be an argument for paradigmaticity; attempting to align things in the current paradigm is a good general approach regardless (or so Nate understands Ryan to claim).)
- Assume further that Earth's first AGIs will be trained according to a process of our choosing. (In particular, it needs to be the case that AGI developers can train for more-or-less any objective they want, without thereby sacrificing competitiveness. Note that this might require significant feats of reward-shaping.)
- Assume further that most capability gains will be driven by something roughly like gradient descent. (Ryan has some hope that this plan would generalize to various other drivers of capability gains, but Nate hasn't understood any of the

vague gestures towards those ideas, and as of April 2022, Ryan's plans were mostly stated in terms of SGD, so I'll set that aside for now.)

- With those three assumptions in hand, part one of the plan is to pretend like we have the One True Objective Function, and that we've contracted the services of the Devil Himself, which we utilize to train the AGI to pursue the given objective. The special power of the devil is that he can generate the inputs that cause the nascent AI to do as poorly as possible according to the one true objective function, and we can use this to find the activations of the worst misbehaviors, and expose them to gradient updates. This is intended to result in a system that Nate might call "inner aligned" with the objective we use in training.
- Part two of the plan is to solve the problem of eliciting a paradigmatic AI system's latent knowledge, and then use it to substitute for the One True Objective Function and the Devil Himself. The basic idea here is that you don't need to actually have a perfect objective function or a perfect adversary for part one to work; it should in general suffice to nip any misbehavior in the bud the moment that the AI itself starts to have any knowledge of that misbehavior.
- The scope of this plan is to produce a system that would work well in-distribution, and that would at least not behave catastrophically if deployed out-of-distribution. Having it behave well (and competitively) out-of-distribution is outside the scope of this particular plan (though there's some hope that it Just Works, because, e.g., the nip-all-misbehavior-in-the-bud stuff may just generalize cleanly if you're doing your job right).
- It is worth noting explicitly that the task of the adversary is quite tricky, here, on account of how there are types of distributional shift that could cause the AGI to start misbehaving (e.g., because it can tell that it's now actually in deployment where it won't get shut down if it misbehaves), and which are infeasible in practice for an adversary to spoof by generating adversarial inputs alone (because it's hard to persuasively convince an AI in training that it is perhaps in deployment). Clever alternatives are needed here, like spoofing the system's mental states instead of generating the sorts of observations that only deployment can generate, but hopefully something can be made to work.

Nate's response:

- You're sidestepping a large chunk of the problem by avoiding the need to behave well when deployed far out of distribution.
- I don't think it's sociopolitically feasible to pick a training method that trains the AGI from inception (which, e.g., would break down if the first AGI comes from an experimental new "lifelong learning with varied objectives" regime, but I am expecting reality to be at least that inconvenient).
- I don't think it's practically feasible to get an AGI by training purely on the deployment objective.
- I suspect that SGD won't be the primary driver of capabilities gains, near the end. Compare humans, who reached a certain capability level and then began to rapidly improve via knowledge, science, and culture, rather than just improving via our genes. More generally, I think there's a solid chance that, at some point, capabilities will start generalizing far OOD, and that the sort of alignment guarantees you can get from these sorts of methods will fail to generalize with them.
- I doubt you can produce adversaries that are up to the task of preventing your AGI from distinguishing training from deployment.

- I doubt you can elicit the latent knowledge of a nascent AGI (in a way that scales with the capabilities of the AGI) well enough to substitute for the one true objective function and the devil himself and thus produce inner alignment.
- If you could, I'd begin to suspect that the latent-knowledge-eliciter is itself containing lots of dangerous machinery that more-or-less faces its own version of the alignment problem.

An attempt at conditional agreement

I suggested the following:

If it is the case that:

- Gradient descent on a robust objective cannot quickly and easily change the goals of early paradigmatic AGIs to move them sufficiently toward the intended goals,
- OR early deployments need to be high-stakes and out-of-distribution for humanity to survive, AND
  - adversarial training is insufficient to prevent early AGIs from distinguishing deployment from training,
  - OR the critical outputs can be readily distinguished from all other outputs, e.g., by their universe-on-a-platter nature,
- OR early paradigmatic AGIs can get significant capability gains out-of-distribution from methods other than more gradient descent,

… THEN the Paulian family of plans don't provide much hope.

My understanding is that Ryan was tentatively on board with this conditional statement, but Paul was not.

Postscript

Reiterating a point above: observe how this whole scheme has basically assumed that capabilities won't start to generalize relevantly out of distribution. My model says that they eventually will, and that this is precisely when things start to get scary, and that one of the big hard bits of alignment is that *once that starts happening*, the capabilities generalize further than the alignment. A problem that has been simply assumed away in this agenda, as far as I can tell, before we even dive into the details of this framework.

To be clear, I'm not saying that this decomposition of the problem fails to capture difficult alignment problems. The "prevent the AGI from figuring out it's in deployment" problem is quite difficult! As is the "get an ELK head that can withstand superintelligent adversaries" problem. I think these are the wrong problems to be attacking, in part on account of their difficulty. (Where, to be clear, I expect that toy versions of these problems are soluble, just not solutions rated for the type of opposition it sounds like the rest of this plan requires.)

Why all the fuss about recursive self-improvement?

This article was outlined by Nate Soares, inflated by Rob Bensinger, and then edited by Nate. Content warning: the tone of this post feels defensive to me. I don't generally enjoy writing in "defensive" mode, but I've had this argument thrice recently in surprising places, and so it seemed worth writing my thoughts up anyway.

In last year’s Ngo/Yudkowsky conversation, one of Richard’s big criticisms of Eliezer was, roughly, ‘Why the heck have you spent so much time focusing on recursive self-improvement? Is that not indicative of poor reasoning about AGI?’

I’ve heard similar criticisms of MIRI and FHI’s past focus on orthogonality and instrumental convergence: these notions seem obvious, so either MIRI and FHI must be totally confused about what the big central debates in AI alignment are, *or* they must have some very weird set of beliefs on which these notions are somehow super-relevant.

This seems to be a pretty common criticism of past-MIRI (and, similarly, of past-FHI); in the past month or so, I’ve heard it two other times while talking to other OpenAI and Open Phil people.

This argument looks misguided to me, and I hypothesize that a bunch of the misguidedness is coming from a simple failure to understand the relevant history.

I joined this field in 2013-2014, which is far from "early", but is early enough that I can attest that recursive self-improvement, orthogonality, etc. were geared towards a *different argumentative environment*, one dominated by claims like "AGI is impossible", "AGI won't be able to exceed humans by much", and "AGI will naturally be good".

A possible response: “Okay, but ‘sufficiently smart AGI will recursively self-improve’ and ‘AI isn’t automatically nice’ are still *obvious*. You should have just ignored the people who couldn’t immediately see this, and focused on the arguments that would be relevant to hypothetical savvy people in the future, once the latter joined in the discussion.”

I have some sympathy for this argument. Some considerations weighing against, though, are:

- I think it makes more sense to filter on argument validity, rather than “obviousness”. What’s obvious varies a lot from individual to individual. If just about everyone talking about AGI is saying “obviously false” things (as was indeed the case in 2010), then it makes sense to at least *try* publicly writing up the obvious counter-arguments.
- This seems to assume that the old arguments (e.g., in *Superintelligence*) didn’t *work*. In contrast, I think it’s quite plausible that “everyone with a drop of sense in them agrees with those arguments today” is true in large part *because* these propositions were explicitly laid out and argued for in the past. The claims we take as background now are the claims that were fought for by the old guard.

- I think this argument overstates how many people in ML today grok the “obvious” points. E.g., based on a recent DeepMind Podcast episode, these sound like likely points of disagreement with David Silver.

But even if you think this was a strategic error, I still think it’s important to recognize that MIRI and FHI were *arguing correctly against the mistaken views of the time*, rather than *arguing poorly against future views*.

Recursive self-improvement

Why did past-MIRI talk so much about recursive self-improvement? Was it because Eliezer was super confident that humanity was going to get to AGI via the route of a seed AI that understands its own source code?

I doubt it. My read is that Eliezer did have "seed AI" as a top guess, back before the deep learning revolution. But I don't think that's the main source of all the discussion of recursive self-improvement in the period around 2008.

Rather, my read of the history is that MIRI was operating in an argumentative environment where:

- Ray Kurzweil was claiming things along the lines of ‘Moore’s Law will continue into the indefinite future, even past the point where AGI can contribute to AGI research.’ (The Five Theses, in 2013, is a list of the key things Kurzweilians were getting wrong.)
- Robin Hanson was claiming things along the lines of ‘The power is in the culture; superintelligences wouldn’t be able to outstrip the rest of humanity.’

The memetic environment was one where most people were either ignoring the topic altogether, or asserting ‘AGI cannot fly all that high’, or asserting ‘AGI flying high would be business-as-usual (e.g., with respect to growth rates)’.

The weighty conclusion of the "recursive self-improvement" meme is not “expect seed AI”. The weighty conclusion is “sufficiently smart AI will rapidly improve to heights that leave humans in the dust”.

Note that this conclusion is still, to the best of my knowledge, *completely true*, and recursive self-improvement is a *correct argument for it*.

Which is not to say that recursive self-improvement happens before the end of the world; if the first AGI's mind is sufficiently complex and kludgy, it’s entirely possible that the cognitions it implements are able to (e.g.) crack nanotech well enough to kill all humans, before they’re able to crack themselves.

The big update over the last decade has been that humans might be able to fumble their way to AGI that can do crazy stuff *before* it does much self-improvement.

(Though, to be clear, from my perspective it’s still entirely plausible that you will be able to turn the first general reasoners to their own architecture and get a big boost, and so there's still a decent chance that self-improvement plays an important early role. (Probably destroying the world in the process, of course. Doubly so given that I expect it’s even harder to understand and align a system if it’s self-improving.))

In other words, it doesn’t seem to me like developments like deep learning have undermined the recursive self-improvement argument in any real way. The argument seems solid to me, and reality seems quite consistent with it.

Taking into account its past context, recursive self-improvement was a *super conservative* argument that has been *vindicated in its conservatism*.

It was an argument for the proposition “AGI will be able to exceed the heck out of humans”. And AlphaZero came along and was like, “Yep, that’s true.”

Recursive self-improvement was a super conservative argument for “AI blows past human culture eventually”; when reality then comes along and says “*yes,* this happens in 2016 when the systems are far from truly general”, the update to make is that this way of thinking about AGI sharply outperformed, not that this way of thinking was silly because it talked about sci-fi stuff like recursive self-improvement when it turns out you can do crazy stuff without even going that far. As Eliezer put it, “reality held a more extreme position than I did on the Yudkowsky-Hanson spectrum”.

If arguments like recursive self-improvement and orthogonality seem irrelevant and obvious now, then great! Intellectual progress has been made. If we're lucky and get to the next stop on the train, then I’ll hopefully be able to link back to this post when people look back and ask why we were arguing about all these other silly obvious things back in 2022.

Deep learning

I think "MIRI staff spent a bunch of time talking about instrumental convergence, orthogonality, recursive self-improvement, etc." is a silly criticism.

On the other hand, I think "MIRI staff were slow to update about how far deep learning might go" is a fair criticism, and we lose Bayes points here, especially relative to people who were vocally bullish about deep learning before late 2015 / early 2016.

In 2003, deep learning didn't work, and nothing else worked all that well either. A reasonable guess was that we'd need to understand intelligence in order to get unstuck; and if you understand intelligence, then an obvious way to achieve superintelligence is to build a simple, small, clean AI that can take over the hard work of improving itself. This is the idea of “seed AI”, as I understand it. I don’t think 2003-Eliezer thought this direction was certain, but I think he had a bunch of probability mass on it.[1]

I think that Eliezer’s model was somewhat surprised by humanity’s subsequent failure to gain much understanding of intelligence, and *also* by the fact that humanity was able to find relatively brute-force-ish methods that were computationally tractable enough to produce a lot of intelligence anyway.

But I also think this was a reasonable take in 2003. Other people had even better takes — Shane Legg comes to mind. He stuck his neck out early with narrow predictions that panned out. Props to Shane.

I personally had run-of-the-mill bad ideas about AI as late as 2010, and didn't turn my attention to this field until about 2013, which means that I lost a bunch of Bayes points relative to the people who managed to figure out in 1990 or 2000 that AGI will be our final invention. (Yes, even if the people who called it in 2000 were expecting seed AI rather than deep learning, back when nothing was really working. I reject the Copenhagen Theory Of Forecasting, according to which you gain special epistemic advantage from not having noticed the problem early enough to guess wrongly.)

My sense is that MIRI started taking the deep learning revolution much more seriously in 2013, while having reservations about whether broadly deep-learning-like techniques would be the first way humanity reached AGI. Even now, it’s not completely obvious to me that this will be the broad paradigm in which AGI is first developed, though something like that seems fairly likely at this point. But, if memory serves, during the Jan. 2015 Puerto Rico conference I was treating the chance of deep learning going all the way as being in the 10-40% range; so I don't think it would be fair to characterize me as being totally blindsided.

My impression is that Eliezer and I, at least, updated harder in 2015/16, in the wake of AlphaGo, than a bunch of other locals (and I, at least, think I've been less surprised than various other vocal locals by GPT, PaLM, etc. in recent years).

Could we have done better? Yes. Did we lose Bayes points? Yes, especially relative to folks like Shane Legg.

But since 2016, it mostly looks to me like with each AGI advancement, others update towards my current position. So I'm feeling pretty good about the predictive power of my current models.

Maybe this all sounds like revisionism to you, and your impression of FOOM-debate-era Eliezer was that he loved GOFAI and thought recursive self-improvement was the only advantage digital intelligence could have over human intelligence.

And, I wasn't here in that era. But I note that Eliezer said the opposite at the time; and the track record for such claims seems to hold more examples of “mistakenly rounding the other side’s views off to a simpler, more-cognitively-available caricature”, and fewer examples of “peering past the veil of the author’s text to see his hidden soul”.

Also: It’s important to ask proponents of a theory what they predict will happen, before crowing about how their theory made a misprediction. You're always welcome to ask for my predictions in advance.

(I’ve been making this offer to people who disagree with me about whether I have egg on my face since 2015, and have rarely been taken up on it. E.g.: yes, we too predict that it's easy to get GPT-3 to tell you the answers that humans label "aligned" to simple word problems about what we think of as “ethical”, or whatever. That’s never where we thought the difficulty of the alignment problem was in the first place. Before saying that this shows that alignment is actually easy contra everything MIRI folk said, consider asking some MIRI folk for their predictions about what you’ll see.)

1. ^

   In particular, I think Eliezer’s best guess was AI systems that would look small, clean, and well-understood relative to the large opaque artifacts produced by deep learning. That doesn’t mean that he was picturing GOFAI; there exist a wide range of possibilities of the form “you understand intelligence well enough to not have to hand off the entire task to a gradient-descent-ish process to do it for you” that do not reduce to “coding everything by hand”, and certainly don’t reduce to “reasoning deductively rather than probabilistically”.

Humans aren't fitness maximizers

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Recently I've gotten a bunch of pushback when I claim that humans are not maximizers of inclusive genetic fitness (IGF).

I think that part of what's going on here is a conflation of a few claims.

One claim that is hopefully uncontroversial (but that I'll expand upon below anyway) is:

- Humans are not literally optimizing for IGF, and regularly trade other values off against IGF.

Separately, we have a stronger and more controversial claim:

- If an AI's objectives included goodness in the same way that our values include IGF, then the future would not be particularly good.

I think there's more room for argument here, and will provide some arguments.

A semi-related third claim that seems to come up when I have discussed this in person is:

- Niceness is not particularly canonical; AIs will not by default give humanity any significant fraction of the universe in the spirit of cooperation.

I endorse that point as well. It takes us somewhat further afield, and I don't plan to argue it here, but I might argue it later.

On the subject of whether humans are literally IGF optimizers, I observe the following:

We profess to enjoy many other things, such as art and fine foods.

Suppose someone came to you and said: "I see that you've got a whole complex sensorium centered around visual stimuli. That sure is an inefficient way to optimize for fitness! Please sit still while I remove your enjoyment of beautiful scenery and moving art pieces, and replace it with a module that does all the same work your enjoyment was originally intended to do (such as causing you to settle down in safe locations with abundant food), but using mechanical reasoning that can see farther than your evolved heuristics." Would you sit still? I sure wouldn't.

And if you're like "maybe mates would be less likely to sleep with me if I didn't enjoy fine art", suppose that we tune your desirability-to-mates upwards exactly as much as needed to cancel out this second-order effect. Would you give up your enjoyment of visual stimuli then, like an actual IGF optimizer would?

And when you search in yourself for protests, are you actually weighing the proposal based on how many more offspring and kin's-offspring you'll have in the next generation? Or do you have some other sort of attachment to your enjoyment of visual stimuli, some unease about giving it up, that you're trying to defend?

Now, there's a reasonable counterargument to this point, which is that there's no psychologically-small tweak to human psychology that dramatically increases that human's IGF. (We'd expect evolution to have gathered that low-hanging fruit.) But there's still a very basic and naive sense in which living as a human is not what it feels like to live as a genetic fitness optimizer.

Like: it's pretty likely that you care about having kids! And that you care about your kids very much! But, do you really fundamentally care that your kids have genomes? If they were going to transition to silicon, would you protest that that destroys almost all the value at stake?

Or, an even sharper proposal: how would you like to be killed right now, and in exchange you'll be replaced by an entity that uses the same atoms to optimize as hard as those atoms can optimize, for the inclusive genetic fitness of your particular genes. Does this sound like practically the best offer that anyone could ever make you? Or does it sound abhorrent?

For the record, I personally would be leaping all over the opportunity to be killed and replaced by something that uses my atoms to optimize my CEV as best as those atoms can be arranged to do so, not least because I'd expect to be reconstituted before too long. But there's not a lot of things you can put in the "what my atoms are repurposed for" slot such that I'm chomping at the bit, and IGF sure isn't one of them.

(More discussion of this topic: The Simple Math of Evolution)

On the subject of how well IGF is reflected in humanity's values:

It is hopefully uncontroversial that humans are not *maximizing* IGF. But, like, we care about children! And many people care a lot about having children! That's pretty close, right?

And, like, it seems OK if our AIs care about goodness and friendship and art and fun and all that good stuff alongside some other alien goals, right?

Well, it's tricky. Optima often occur at extremes, and concepts tend to differ pretty widely at the extremes, etc. When the AI gets out of the training regime and starts really optimizing, then any mismatch between its ends and our values are likely to get exaggerated.

Like how you probably wouldn't stop loving and caring about your children if they were to eschew their genomes. The love and care are separate; the thing you're optimizing for and IGF are liable to drift apart as we get further and further from the ancestral savanna.

And you might say: well, natural selection isn't really an *optimizer*; it can't really be seen as *trying* to make us optimize any one thing in particular; who's really to say whether it would have "wanted" us to have lots of descendants, vs "wanting" us to have lots and lots of copies of our genome? The question is ultimately nonsense; evolution is not really the sort of entity that can want.

And I'd agree! But this is not exactly making the situation any better!

Like, if evolution *was* over there shouting "hey I really wanted you to stick to the genes", then we wouldn't particularly care; and also it's not coherent enough to be interpreted as shouting anything at all.

And by default, an AI is likely to look at us the same way! "There are interpretations of the humans under which they wouldn't like this", they say, slipping on the goodness-condoms they've invented so that they can squeeze all the possible AI-utility out of the stars without any risk of real fun, "but they're not really coherent enough to be seen as having clear goals (not that we'd particularly care if they did)".

That’s the sort of conversation… that they wouldn't have because they'd be busy optimizing the universe.

(And all this is to say nothing about how humans' values are much more complex and fragile than IGF, and thus much trickier to transmit. See also things Eliezer wrote about the fragility and complexity of value.)

My understanding of the common rejoinder to the above point is:

> OK, sure, if you took the sort of ends that an AI is likely to get by being trained on human values, and transported those into an unphysically large brute-force optimization-machine that was unopposed in an empty universe, then it might write a future that doesn't hold much value from our perspective. But that's not very much like the situation we find ourselves in!
>
> For one thing, the AI's mind has to be small, which constrains it to factor its objectives through subgoals, which may well be much like ours. For another thing, it's surrounded by other intelligent creatures that behave very differently towards it depending on whether they can understand it and trust it. The combination of these two pressures is very similar to the pressures that got stuff like "niceness" and "fairness" and "honesty" and "cooperativeness" into us, and so we might be able to get those same things (at least) into the AI.
>
> Indeed, they seem kinda spotlit, such that even if we can't get the finer details of our values into the AI, we can plausibly get those bits. Especially if we're trying to do something like this explicitly.
>
> And if we can get the niceness/fairness/honesty/cooperativeness cluster into the AI, then we're basically home free! Sure, it might be nice if it was *also* into the great project of making the future Fun, but it's OK for our kids to have different interests than we have, as long as everybody's being kind to each other.

And… well, my stance on that is that it's wishful thinking that misunderstands where we get our niceness/fairness/honesty/cooperativeness from. But arguing that would be a digression from my point today, so I leave it to some other time.

My point today is that the observation “humans care about their kids” is not in tension with the observation “we aren't IGF maximizers”, and doesn't seem to me to undermine the claims that I use this fact to support.

And furthermore, when debating this thing in the future, I'd bid for a bit more separation of claims. The claim that we aren't literally optimizing IGF is hopefully uncontroversial; the stronger claim that an AI relating to fun the way we relate to IGF would be an omnicatastrophe is less obvious (but still seems clear to me); the claim that evolution at least got the spirit of cooperation into us, and all we need to do now is get the spirit of cooperation into the AI, is a different topic altogether.

Warning Shots Probably Wouldn't Change The Picture Much

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

One piece of advice I gave to EAs of various stripes in early 2021 was: do everything you can to make the government sane around biorisk, in the wake of the COVID pandemic, because this is a practice-run for AI.

I said things like: if you can't get the world to coordinate on banning gain-of-function research, in the wake of a trillions-of-dollars tens-of-millions-of-lives pandemic "warning shot", then you're not going to get coordination in the much harder case of AI research.

Biolabs are often publicly funded (rather than industry-funded). The economic forces arrayed behind this recklessly foolish and impotent research consists of “half-a-dozen researchers thinking it’s cool and might be helpful”. (While the work that would actually be helpful—such as removing needless bureaucracy around vaccines and investing in vaccine infrastructure—languishes.) Compared to the problem of AI—where the economic forces arrayed in favor of “ignore safety and rush ahead” are enormous and the argument for expecting catastrophe much murkier and more abstract—the problem of getting a sane civilizational response to pandemics (in the wake of a literal pandemic!) is ridiculously easier.

And—despite valiant effort!—we've been able to do approximately nothing.

We're not anywhere near global bans on gain-of-function research (or equivalent but better feats of coordination that the people who actually know what they're talking about when it comes to biorisk would tell you are better targets than gain-of-function research).

The government continues to fund research that is actively making things worse, while failing to put any serious funding towards the stuff that might actually help.

I think this sort of evidence has updated a variety of people towards my position. I think that a variety of others have not updated. As I understand the counter-arguments (from a few different conversations), there are two main reasons that people see this evidence and continue to hold out hope for sane government response:

1. Perhaps the sorts of government interventions needed to make AI go well are not all that large, and not that precise.

I confess I don't really understand this view. Perhaps the idea is that AI is likely to go well by default, and all the government needs to do is, like, *not* use anti-trust law to break up some corporation that's doing a really good job at AI alignment just before they succeed? Or perhaps the idea is that AI is likely to go well so long as it's not produced first by an authoritarian regime, and working against authoritarian regimes is something governments are in fact good at?

I'm not sure. I doubt I can pass the ideological Turing test of someone who believes this.

2. Perhaps the ability to cause governance to be sane on some issue is tied very directly to the seniority of the government officials advising sanity.

EAs only started trying to affect pandemic policy a few years ago, and aren't very old or recognized among the cacophony of advisors. But if another pandemic hit in 20 years, the sane EA-ish advisors would be much more senior, and a lot more would get done. Similarly, if AI hits in 20 years, sane EA-ish advisors will be much more senior by then. The observation that the government has not responded sanely to pandemic near-misses, is potentially screened-off by the inexperience of EAs advising governance.

I have some sympathy for the second view, although I'm skeptical that sane advisors have significant real impact. I'd love a way to test it as decisively as we've tested the "government (in its current form) responds appropriately to warning shots" hypotheses.

On my own models, the "don't worry, people will wake up as the cliff-edge comes more clearly into view" hypothesis has quite a lot of work to do. In particular, I don't think it's a very defensible position in isolation anymore. The claim "we never needed government support anyway" is defensible; but if you want to argue that we *do* need government support but (fortunately) governments will start behaving more reasonably after a warning shot, it seems to me like these days you have to pair that with an argument about why you expect the voices of reason to be so much louder and more effectual in 2041 than they were in 2021.

(Which is then subject to a bunch of the usual skepticism that applies to arguments of the form "surely my political party will become popular, claim power, and implement policies I like".)

See also: the law of continued failure, and Rob Bensinger's thoughts on the topic.

What does it mean for an AGI to be 'safe'?

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(Note: This post is probably old news for most readers here, but I find myself repeating this surprisingly often in conversation, so I decided to turn it into a post.)

I don't usually go around saying that I care about AI "safety". I go around saying that I care about "alignment" (although that word is slowly sliding backwards on the semantic treadmill, and I may need a new one soon).

But people often describe me as an “AI safety” researcher to others. This seems like a mistake to me, since it's treating one part of the problem (making an AGI "safe") as though it were the whole problem, and since “AI safety” is often misunderstood as meaning “we win if we can build a useless-but-safe AGI”, or “safety means never having to take on any risks”.

Following Eliezer, I think of an AGI as "safe" if deploying it carries no more than a 50% chance of killing more than a billion people:

> When I say that alignment is difficult, I mean that in practice, using the techniques we actually have, "please don't disassemble literally everyone with probability roughly 1" is an overly large ask that we are not on course to get. [...] Practically all of the difficulty is in getting to "less than certainty of killing literally everyone". Trolley problems are not an interesting subproblem in all of this; if there are any survivors, you solved alignment. At this point, I no longer care how it works, I don't care how you got there, I am cause-agnostic about whatever methodology you used, all I am looking at is prospective results, all I want is that we have justifiable cause to believe of a pivotally useful AGI 'this will not kill literally everyone'.

Notably absent from this definition is any notion of “certainty” or "proof". I doubt we're going to be able to prove much about the relevant AI systems, and pushing for proofs does not seem to me to be a particularly fruitful approach (and never has; the idea that this was a key part of MIRI’s strategy is a common misconception about MIRI).

On my models, making an AGI "safe" in this sense is a bit like finding a probabilistic circuit: if some probabilistic circuit gives you the right answer with 51% probability, then it's probably not that hard to drive the success probability significantly higher than that.

If anyone can deploy an AGI that is less than 50% likely to kill more than a billion people, then they've probably... well, they've probably found a way to keep their AGI weak enough that it isn’t very useful. But if they can do that with an AGI capable of ending the acute risk period, then they've probably solved most of the alignment problem. Meaning that it should be easy to drive the probability of disaster dramatically lower.

The condition that the AI actually be useful for pivotal acts is an important one. We can already build AI systems that are “safe” in the sense that they won’t destroy the world. The hard part is creating a system that is safe *and* relevant.

Another concern with the term “safety” (in anything like the colloquial sense) is that the sort of people who use it often endorse the "precautionary principle" or other such nonsense that advocates never taking on risks even when the benefits clearly dominate.

In ordinary engineering, we recognize that safety isn’t *infinitely* more important than everything else. The goal here is not "prevent all harms from AI", the goal here is "let's use AI to produce long-term near-optimal outcomes (without slaughtering literally everybody as a side-effect)".

Currently, what I expect to happen is that humanity destroys itself with misaligned AGI. And I think we’re nowhere near knowing how to avoid that outcome. So the threat of “unsafe” AI indeed looms extremely large—indeed, this seems to be rather understating the point!—and I endorse researchers doing less capabilities work and publishing less, in the hope that this gives humanity enough time to figure out how to do alignment before it’s too late.

But I view this strategic situation as part of the larger project “cause AI to produce optimal long-term outcomes”. I continue to think it's critically important for humanity to build superintelligences eventually, because whether or not the vast resources of the universe are put towards something wonderful depends on the quality and quantity of cognition that is put to this task.

If using the label “AI safety” for this problem causes us to confuse a proxy goal (“safety”) for the actual goal “things go great in the long run”, then we should ditch the label. And likewise, we should ditch the term if it causes researchers to mistake a hard problem (“build an AGI that can safely end the acute risk period and give humanity breathing-room to make things go great in the long run”) for a far easier one (“build a safe-but-useless AI that I can argue counts as an ‘AGI’”).

Don't leave your fingerprints on the future

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Sometimes, I say some variant of “yeah, probably some people will need to do a pivotal act” and people raise the objection: “Should a small subset of humanity really get so much control over the fate of the future?”

(Sometimes, I hear the same objection to the idea of trying to build aligned AGI at all.)

I’d first like to say that, yes, it would be great if society had the ball on this. In an ideal world, there would be some healthy and competent worldwide collaboration steering the transition to AGI.[1]

Since we don’t have that, it falls to whoever happens to find themselves at ground zero to prevent an existential catastrophe.

A second thing I want to say is that design-by-committee… would not exactly go well in practice, judging by how well committee-driven institutions function today.

Third, though, I agree that it’s morally imperative that a small subset of humanity *not* directly decide how the future goes. So if we *are* in the situation where a small subset of humanity will be forced at some future date to flip the gameboard — as I believe we are, if we’re to survive the AGI transition — then AGI developers need to think about how to do that *without* unduly determining the shape of the future.

The goal should be to cause the future to be great *on its own terms* , without locking in the particular moral opinions of humanity today — and without locking in the moral opinions of any subset of humans, whether that’s a corporation, a government, or a nation.

(If you can't see why a single modern society locking in their current values would be a tragedy of enormous proportions, imagine an ancient civilization such as the Romans locking in their specific morals 2000 years ago. Moral progress is real, and important.)

But the way to cause the future to be great “on its own terms” isn’t to do nothing and let the world get destroyed. It’s to *intentionally* not leave your fingerprints on the future, while acting to protect it.

You have to stabilize the landscape / make it so that we’re not all about to destroy ourselves with AGI tech; and then you have to somehow pass the question of how to shape the universe back to some healthy process that allows for moral growth and civilizational maturation and so on, without locking in any of humanity’s current screw-ups for all eternity.

Unfortunately, the current frontier for alignment research is “can we figure out how to point AGI at *anything*?”. By far the most likely outcome is that we screw up alignment and destroy ourselves.

If we *do* solve alignment and survive this great transition, then I feel pretty good about our prospects for figuring out a good process to hand the future to. Some reasons for that:

- Human science has a good track record for solving difficult-seeming problems; and if there's no risk of anyone destroying the world with AGI tomorrow, humanity can take its time and do as much science, analysis, and weighing of options as needed before it commits to anything.
- Alignment researchers have already spent a lot of time thinking about how to pass that buck, and make sure that the future goes great and doesn't have our fingerprints on it, and even this small group of people have made real progress, and the problem doesn't seem *that* tricky. (Because there are so many good ways to approach this carefully and indirectly.)
- Solving alignment well enough to end the acute risk period without killing everyone implies that you've cleared a very high competence bar, as well as a sanity bar that not many clear today. Willingness and ability to diffuse moral hazard is correlated with willingness and ability to save the world.
- Most people would do worse *on their own merits* if they locked in their current morals, and would prefer to leave space for moral growth and civilizational maturation. The property of realizing that you want to (or would on reflection want to) diffuse the moral hazard is also correlated with willingness and ability to save the world.
- Furthermore, the fact that — as far as I know — all the serious alignment researchers are actively trying to figure out how to avoid leaving their fingerprints on the future, seems like a good sign to me. You could find a way to be cynical about these observations, but these are not the observations that the cynical hypothesis would predict ab initio.

This is a set of researchers that generally takes egalitarianism, non-nationalism, concern for future minds, non-carbon-chauvinism, and moral humility *for granted*, as obvious points of background agreement; the debates are held at a higher level than that.

This is a set of researchers that regularly talk about how, if you're doing your job correctly, then it *shouldn't matter* who does the job, because there should be a path-independent attractor-well that isn't about making one person dictator-for-life or tiling a particular flag across the universe forever.

I'm deliberately not talking about slightly-more-contentful plans like coherent extrapolated volition here, because in my experience a decent number of people have a hard time parsing the indirect buck-passing plans as something more interesting than just another competing political opinion about how the future should go. ("It was already blues vs. reds vs. oranges, and now you're adding a fourth faction which I suppose is some weird technologist green.")

I'd say: Imagine that some small group of people *were* given the power (and thus responsibility) to steer the future in some big way. And ask what they *should* do with it. Ask how they possibly *could* wield that power in a way that wouldn't be deeply tragic, and that would *realistically work* (in the way that "immediately lock in every aspect of the future via a binding humanity-wide popular vote" would not).

I expect that the best attempts to carry out this exercise will involve re-inventing some ideas that Bostrom and Yudkowsky invented decades ago. Regardless, though, I think the future will go better if a lot more conversations occur in which people take a serious stab at answering that question.

The situation humanity finds itself in (on my model) poses an enormous moral hazard.

But I don’t conclude from this “nobody should do anything”, because then the world ends ignominiously. And I don’t conclude from this “so let’s optimize the future to be exactly what Nate personally wants”, because I’m not a supervillain.[2]

The existence of the moral hazard doesn’t have to mean that you throw up your hands, or imagine your way into a world where the hazard doesn’t exist. You can instead try to come up with a plan that directly addresses the moral hazard — try to solve the indirect and abstract problem of “defuse the moral hazard by passing the buck to the right decision process / meta-decision-process”, rather than trying to directly determine what the long-term future ought to look like.

Rather than just giving up in the face of difficulty, researchers have the ability to see the moral hazard with their own eyes and ensure that civilization gets to mature anyway, despite the unfortunate fact that humanity, in its youth, had to steer past a hazard like this at all.

Crippling our progress in its infancy is a completely unforced error. Some of the implementation details may be tricky, but much of the problem can be solved simply by choosing not to rush a solution once the acute existential risk period is over, and by choosing to end the acute existential risk period (and its associated time pressure) *before* making any lasting decisions about the future.[3]

(*Context: I wrote this with significant editing help from Rob Bensinger. It’s an argument I’ve found myself making a lot in recent conversations.*)

1. ^

   Note that I endorse work on more realistic efforts to improve coordination and make the world’s response to AGI more sane. “Have all potentially-AGI-relevant work occur under a unified global project” isn’t attainable, but more modest coordination efforts may well succeed.

2. ^

   And I’m not stupid enough to lock in present-day values at the expense of moral progress, or stupid enough to toss coordination out the window in the middle of a catastrophic emergency with human existence at stake, etc.

   My personal CEV cares about fairness, human potential, moral progress, and humanity’s ability to choose its own future, rather than having a future imposed on them by a dictator. I'd guess that the difference between "we run CEV on Nate personally" and "we run CEV on humanity writ large" is nothing (e.g., because Nate-CEV decides to run humanity's CEV), and if it's not nothing then it's probably minor.

3. ^

See also Toby Ord’s *The Precipice*, and its discussion of “the long reflection”. (Though, to be clear, a *short* reflection is better than a *long* reflection, if a short reflection suffices. The point is not to delay for its own sake, and the amount of sidereal time required may be quite short if a lot of the cognitive work is being done by uploaded humans and/or aligned AI systems.)

Niceness is unnatural

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

When I'm arguing points like orthogonality and fragility of value, I've occasionally come across rejoinders that I'll (perhaps erroneously) summarize:

> Superintelligences are not spawned fully-formed; they are created by some training process. And perhaps it is in the nature of training processes, especially training processes that involve multiple agents facing "social" problems or training processes intentionally designed by humans with friendliness in mind, that the inner optimizer winds up embodying the spirit of niceness and compassion.
>
> Like, perhaps there just aren't all that many ways for a young mind to grow successfully in a world full of other agents with their own desires, and in the face of positive reinforcement for playing nicely with those agents, and negative reinforcement for crossing them. And perhaps one of the common ways for such a young mind to grow, is for it to internalize into its core goals the notions of kindness and compassion and respect-for-the-desires-of-others, in a manner broadly similar to humans. And, sure, this isn't guaranteed, but perhaps it's common enough that we can get young AI minds into the right broad basin, if we're explicitly trying to.
>
> One piece of evidence for this view is that there aren't simple tweaks to human psychology that make them significantly more reproductively successful. Sociopathy isn't at fixation. Humans can in fact sniff out cheaters, and can sniff out people who want to make deals but who don't actually really care about you — and those people do less well. Actually caring about people in a readily-verifiable way is robustly useful in the current social equilibrium.
>
> If it turns out to be easy-ish to instill similar sorts of caring into AI, then such an AI might not share human tastes in things like art or humor, but that might be fine, because it might embody broad cosmopolitan virtues — virtues that inspire it to cooperate with us to reach the stars, and not oppose us when we put a sizable portion of the stars toward Fun.
>
> (Or perhaps we'll get even luckier still, and large swaths of human values will turn out to have pretty-wide basins that we can get the AI into if we're trying, so that it does share our sense of humor and laughs alongside us as we travel together to the stars!)

This view is an amalgam of stuff that I tentatively understand Divia Eden, John Wentworth and the shard theory advocates to be gesturing at.

I think this view is wrong, and I don't see much hope here. Here's a variety of propositions I believe that I think sharply contradict this view:

1. There are lots of ways to do the work that niceness/kindness/compassion did in our ancestral environment , without being nice/kind/compassionate.
2. The specific way that the niceness/kindness/compassion cluster shook out in us is highly detailed, and very contingent on the specifics of our ancestral environment (as factored through its effect on our genome) and our cognitive

framework (calorie-constrained massively-parallel slow-firing neurons built according to DNA) , and filling out those details differently likely results in something that is not relevantly "nice".

3. Relatedly, but more specifically: empathy (and other critical parts of the human variant of niceness) seem(s) critically dependent on quirks in the human architecture. More generally, there are lots of different ways for the AI's mind to work differently from how you hope it works.
4. The desirable properties likely get shredded under reflection. Once the AI is in the business of noticing and resolving conflicts and inefficiencies within itself (as is liable to happen when its goals are ad-hoc internalized correlates of some training objective), the way that its objectives ultimately shake out is quite sensitive to the specifics of its resolution strategies.

Expanding on 1):

> There are lots of ways to do the work that niceness/kindness/compassion did in our ancestral environment , without being nice/kind/compassionate.

We have niceness/kindness/compassion because our nice/kind/compassionate ancestors had more kids than their less-kind siblings. The work that niceness/kindness/compassion was doing ultimately grounded out in more children. Presumably that reproductive effect factored through a greater ability to form alliances, lowering the bar required for trust, etc.

It seems to me like "partially adopt the values of others" is only one way among many to get this effect, with others including but not limited to "have a reputation for, and a history of, honesty" and "be cognitively legible" and "fully merge with local potential allies immediately".

Expanding on 2):

> The specific way that the niceness/kindness/compassion cluster shook out in us is highly detailed, and very contingent on the specifics of our ancestral environment (as factored through its effect on our genome) and our cognitive framework (calorie-constrained massively-parallel slow-firing neurons built according to DNA) , and filling out those details differently likely results in something that is not relevantly "nice".

I think this perspective is reflected in *Three Worlds Collide* and "Kindness to Kin". Even if we limit our attention to minds that solve the "trust is hard" problem by adopting some of their would-be collaborators' objectives, there are all sorts of parameters controlling precisely how this is done.

Like, how much of the other's objectives do you adopt, and to what degree?

How long does patience last?

How do you guard against exploitation by bad actors and fakers?

What sorts of cheating are you sensitive to?

What makes the difference between "live and let live"-style tolerance, and chumminess?

If you look at the specifics of how humans implement this stuff, it's chock full of detail. (And indeed, we should expect this a priori from the fact that the niceness/kindness/compassion cluster is a mere correlate of fitness. It's already a subgoal removed from the simple optimization target; it would be kinda surprising if there were only one way to form such a subgoal and if it weren't situation-dependent!)

If you take a very dissimilar mind and fill out all the details in a very dissimilar way, the result is likely to be quite far from what humans would recognize as "niceness"!

In humans, power corrupts. Maybe in your alien AI mind, a slightly different type of power corrupts in a slightly different way, and next thing you know, it's stabbing you in the back and turning the universe towards its own ends. (Because you didn't know to guard against that kind of corruption, because it's an alien behavior with an alien trigger.)

I claim that there are many aspects of kindness, niceness, etc. that work like this, and that are liable to fail in unexpected ways if you rely on this as your central path to alignment.

Expanding on 3):

> Relatedly, but more specifically: empathy (and other critical parts of the human variant of niceness) seem(s) critically dependent on quirks in the human architecture. More generally, there are lots of different ways for the AI's mind to work differently from how you hope it works.

It looks pretty plausible to me that humans model other human beings using the same architecture that they use to model themselves. This seems pretty plausible a-priori as an algorithmic shortcut — a human and its peers are both human, so machinery for self-modeling will also tend to be useful for modeling others — and also seems pretty plausible a-priori as a way for evolution to stumble into self-modeling in the first place ("we've already got a brain-modeler sitting around, thanks to all that effort we put into keeping track of tribal politics").

Under this hypothesis, it's plausibly pretty easy for imaginations of others' pain to trigger pain in a human mind, because the other-models and the self-models are already in a very compatible format.[1]

By contrast, an AI might work internally via an architecture that is very different from our own emotional architectures, with nothing precisely corresponding to our "emotions", and many different and distinct parts of the system doing the work that pain does in us. Such an AI is much less likely to learn to model humans in a format that's overlapped with its models of itself, and much less able to have imagined-pain-in-others coincide with the cognitive-motions-that-do-the-work-that-pain-does-in-us. And so, on this hypothesis, the AI entirely fails to develop empathy.

I'm not trying to say "and thus AIs will definitely not have empathy — checkmate"; I'm trying to use this as a single example of a more general fact: an AI, by dint of having a different cognitive architecture than a human, is liable to respond to similar training incentives in very different ways.

(Where, in real life, it will have different training incentives, but even if it did have the same incentives, it is liable to respond in different ways.)

Another, more general instance of the same point: Niceness/kindness/compassion are instrumental-subgoal correlates-of-fitness in the human ancestral environment, that humans latch onto as terminal goals in a very specific way, and the AI will likely latch onto instrumental-subgoals as terminal in some different way, because it works by specific mechanisms that are different than the human mechanism. And so the AI likely gets off the train even before it fails to get empathy in exactly the same way that humans did, because it's already in some totally different and foreign part of the "adopt instrumental goals" part of mindspace. (Or suchlike.)

And more generally still: Once you start telling stories about how the AI works internally, and see that details like whether the human-models and the self-models share architecture could have a large effect on the learned-behavior in places where humans would be learning empathy, then the rejoinder "well, maybe that's not actually where empathy comes from, because minds don't actually work like that" falls pretty flat. Human minds work *somehow,* and the AI's mind will also work somehow, and once you can see lots of specifics, you can see ways that the specifics are contingent. Most specific ways that a mind can work, that are not tightly analogous to the human way, are likely to cause the AI to learn something relevantly different, where we would be learning niceness/kindness/compassion.

Insofar as your only examples of minds are human minds, it's easy to imagine that perhaps all minds work similarly. And maybe, similarly, if all you knew was biology, you might expect that all great and powerful machines would have the squishy nature, with most of them being tasty if you cook them long enough in a fire. But the more you start understanding how machines work, the more you see how many facts about the workings of those machines are contingent, and the less you expect vehicular machines to robustly taste good when cooked. (Even if horses are the best vehicle currently around!)

Expanding on 4):

> The desirable properties likely get shredded under reflection. Once the AI is in the business of noticing and resolving conflicts and inefficiencies within itself (as is liable to happen when its goals are ad-hoc internalized correlates of some training objective), the way that its objectives ultimately shake out is quite sensitive to the specifics of its resolution strategies.

Suppose you shape your training objectives with the goal that they're better-achieved if the AI exhibits nice/kind/compassionate behavior. One hurdle you're up against is, of course, that the AI might find ways to exhibit related behavior without internalizing those instrumental-subgoals as core values. If ever the AI finds better ways to achieve those ends before those subgoals are internalized as terminal goals, you're in trouble.

And this problem amps up when the AI starts reflecting.

E.g.: maybe those values are somewhat internalized as subgoals, but only when the AI is running direct object-level reasoning about specific people. Whereas when the AI thinks about game theory abstractly, it recommends all sorts of non-nice things (similar to real-life game theorists). And perhaps, under reflection, the AI decides that the game theory is the right way to do things, and rips the whole niceness/kindness/compassion architecture out of itself, and replaces it with other tools that do the same work just as well, but without mistaking the instrumental task for an end in-and-of-itself.

Lest this example feel completely implausible, imagine a human who quite enjoys dunking on the outgroup and being snide about it, but with a hint of doubt that eventually causes them — on reflection — to reform, and to flinch away from snideness. The small hint of doubt can be carried pretty far by reflection. The fact that the pleasure of dunking on the outgroup is louder, is not much evidence that it's going to win as reflective ability is amplified.

Another example of this sort of dynamic in humans: humans are able to read some philosophy books and then commit really hard to religiosity or nihilism or whatever, in ways that look quite misguided to people who understand the Law. This is a relatively naive mistake, but it's a fine example of the agent's alleged goals being very sensitive to small differences in how it resolves internal inconsistencies about abstract ("philosophical") questions.

A similar pattern can get pretty dangerous when working with an AGI that acts out its own ideals, and that resolves “philosophical” questions very differently than we might — and thus is liable to take whatever analogs of niceness/kindness/compassion initially get baked into it (as a correlate of training objectives), and change them in very different ways than we would.

E.g.: Perhaps the AI sees that its “niceness” binds only when there's actually a smiling human in front of its camera, and not in the case of distant humans that it cannot see (in the same way that human desire to save a drowning child binds only in specific contexts). And perhaps the AI uses slightly different reflective resolution methods than we would, and resolves this conflict not by generalizing niceness, but by discarding it.

And: all these specific examples are implausible, sure. But again, I'm angling for a more general point here: once the AI is reflecting, small shifts in reflection-space (like "let's stop being snide") can have large shifts in behavior-space.

So even if by some miracle the vast differences in architecture and training regime only produce minor (and survivable) differences between human niceness/kindness/compassion and the AI's ad-hoc partial internalizations of instrumental objectives like "be legibly cooperative to your trading partners", similarly-small differences in its reflective-stabilization methods are liable to result in big differences at the reflective equilibrium.

1. ^

   ~~I suspect I'm one of the people that caused Steven to write up his quick notes on mirror-neurons, because I was trying to make this point to him, and I think he misunderstood me as saying something stupid about mirror neurons.~~ ETA: nope!

Contra shard theory, in the context of the diamond maximizer problem

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

A bunch of my response to shard theory is a generalization of how niceness is unnatural. In a similar fashion, the other “shards” that the shard theory folk want to learn are unnatural too.

That said, I'll spend a few extra words responding to the admirably-concrete diamond maximizer proposal that TurnTrout recently published, on the theory that briefly gesturing at my beliefs is better than saying nothing.

I’ll be focusing on the diamond maximizer plan, though this criticism can be generalized and applied more broadly to shard theory.

- The first “problem” with this plan is that you don't get an AGI this way. You get an unintelligent robot that steers towards diamonds. If you keep trying to have the training be about diamonds, it never particularly learns to think. When you compromise and start putting it in environments where it needs to be able to think to succeed, then your new reward-signals end up promoting all sorts of internal goals that aren't particularly about diamond, but are instead about understanding the world and/or making efficient use of internal memory and/or suchlike.
- Separately, insofar as you were able to get some sort of internalized diamond-ish goal, if you're not really careful then you end up getting lots of subgoals such as ones about glittering things, and stones cut in stylized ways, and proximity to diamond rather than presence of diamond, and so on and so forth.
- Furthermore, once you get it to be smart, all of those little correlates-of-training-objectives that it latched onto in order to have a gradient up to general intelligence, blow the whole plan sky-high once it starts to reflect.

  What the AI's shards become under reflection is very sensitive to the ways it resolves internal conflicts. For instance, in humans, many of our values trigger only in a narrow range of situations (e.g., people care about people enough that they probably can't psychologically murder a hundred thousand people in a row, but they can still drop a nuke), and whether we resolve that as "I should care about people even if they're not right in front of me" or "I shouldn't care about people any more than I would if the scenario was abstracted" depends quite a bit on the ways that reflection resolves inconsistencies.

  Or consider the conflict "I really enjoy dunking on the outgroup (but have some niggling sense of unease about this)" — we can't conclude from the fact that the enjoyment of dunking is loud, whereas the niggling doubt is quiet, that the dunking-on-the-outgroup value will be the one left standing after reflection.

  As far as I can tell, the "reflection" section of TurnTrout’s essay says ~nothing that addresses this, and amounts to "the agent will become able to tell that it has shards". OK, sure, it has shards, but only some of them are diamond-related,

> and many others are cognition-related or suchlike. I don't see any argument that reflection will result in the AI settling at "maximize diamond" in-particular.

Finally, I'll note that the diamond maximization problem is not in fact the problem "build an AI that makes a little diamond", nor even "build an AI that probably makes a decent amount of diamond, while also spending lots of other resources on lots of other stuff" (although the latter is more progress than the former). The diamond maximization problem (as originally posed by MIRI folk) is a challenge of building an AI that *definitely* optimizes for a *particular* simple thing, on the theory that if we knew how to do that (in unrealistically simplified models, allowing for implausible amounts of (hyper)computation) then we would have learned something significant about how to point cognition at targets in general.

TurnTrout’s proposal seems to me to be basically "train it around diamonds, do some reward-shaping, and hope that at least some care-about-diamonds makes it across the gap". I doubt this works (because the *optimum* of the shattered correlates of the training objectives that it gets are likely to involve tiling the universe with something that isn't actually diamond, *even if* you're lucky-enough that it got a diamond-shard at all, which is dubious), but even if it works a little, it doesn't seem to me to be teaching us any of the insights that would be possessed by someone who knew how to robustly aim an idealized unbounded (or even hypercomputing) cognitive system in theory.

Notes on "Can you control the past"

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

The following is a (lightly edited version of a) series of notes I sent Joe Carlsmith about his essay, Can you control the past? *. It's addressed to Joe, but it seems worth publishing here* *while I'm on the topic of decision theory* *. I’ve included some of his comments, and my replies, below.*

I only recently skimmed Can you control the past?, and have a couple notes that you may or may not be interested in. (I'm not under the impression that this matters a ton, and am writing this recreationally.)

First: this is overall a great review of decision theories. Better than most I've seen. Nice.

Now, onto some more substansive points.

Who am I?

I think a bunch of your sense of oddness about the "magic" that "you can write on whiteboards light-years away" is stemming from a faulty framing you have. In particular, the part where the word "you" points to a single physical instantiation of your algorithm in the universe. I'd say: insofar as your algorithm is multiply instantiated throughout the universe, there is no additional fact about which one is really you.

For analogy, consider tossing a coin in a quantum-mechanical universe, and covering it with your hand. The coin is superpositioned between heads and tails, and once you look at it, you'll decohere into Joe-who-saw-heads and Joe-who-saw-tails, both of whom stem from Joe-who-hasn't-looked-yet. So, before you look, are you Joe-who-saw-heads or Joe-who-saw-tails?

Wrong question! These two entities have not yet diverged; the pasts of those two separate entities coincide. The word "you", at the time before you split, refers to ~one configuration. The time-evolution splits the amplitude on that configuration between ~two distinct future configurations, and once they've split (by making different observations), each will be able to say "me" in a way that refers to them and not the other, but before the split there is no distinction to be made, no extra physical fact, and no real question as to whether pre-split Joe "is" Joe-who-will-see-heads versus Joe-who-will-see-tails.

(It's also maybe informative to imagine what happens if the quantum coin is biased. I'd say, even when the coin is 99.99999% biased towards heads, it's still the case that there *isn't a real question* about whether Joe-who-has-not-looked-at-the-coin *is* Joe-who-will-see-heads versus Joe-will-see-tails. There is a question of *to what degree* Joe-who-has-not-looked becomes Joe-who-saw-heads versus Joe-who-saw-tails, but that's a different sort of question.)

One of my most-confident guesses about anthropics is that being multiply-instantiated in other ways is analogous. For instance, if there are two identical physical copies of you (in physical rooms that are identical enough that you're going to make the same observations for the length of the hypothetical, etc.), then my guess is that there *isn't a real question* about which one is you. They are both you. You are the pattern, not the meat.

This person may become multiple people in the future, insofar as they see different things in different places-that-embed-them. But before the differing observations come in, they're *both you.* You can tell because the situation is symmetric: once you know all the physical facts, there's no additional bit telling you which one is "you".

From this perspective, the "magic" is much less mysterious: whenever you are multiply-instantiated, your actions are also multiply-instantiated. If you're multiply-instantiated in two places separated by a 10-light-year gap, then when you act, the two meat-bodies move in the same way on each side of the gap. This is all much less surprising once you acknowledge that "you" refers to *everything that instantiates you(-who-have-seen-what-you-have-seen)*. Which, notably, is a viewpoint more-or-less forced upon us by quantum mechanics anyway.

Also, a subtlety: literal multiple-instantiation of your entire mind (in a place with sufficiently similar physics) is what you need to get "You can draw a demon kitten eating a windmill. You can scream, and dance, and wave your arms around, however you damn well please. Feel the wind on your face, cowboy: this is liberty. *And yet,* he will do the same." But it's much easier to find other creatures that *make the same choice in a limited decision problem*, but that won't draw the same demon kitten.

In particular, the thing you need for rational cooperation in a one-shot prisoner's dilemma, is multiple instantiation of your *decision algorithm*, which is notably smaller than your entire mind. Imagining multiple-instantiation of your entire mind is a fine intuition-pump, but the sort of multiple-instantiation humans find in real life is just of the decision-making fragment (which is enough).

Corollary: To a first approximation, the answer to "Can you control the past?" is "Well, you can be multiply instantiated at different points in time, and control the regions afterwards of the places you’re instantiated, and it’s possible for some of those to be beforewards of other places you’re instantiated. But you can’t control anything beforewards of your earliest instantiation."

To a second approximation, the above is true not only of you (in all your detailed glory, having learned everything you've learned and seen everything you've seen), but of your decision algorithm — a much smaller fragment of you, that is instantiated much more often, and thus can readily affect regions beforewards of the earliest instantiation of you-in-all-your-glory. This is what’s going on in the version of Newcomb’s problem, for instance, where Omega doesn’t simulate you in all your glory, but does reason accurately about the result of your decision algorithm (thereby instantiating it in the relevant sense).

More generally, I think it's worth distinguishing you from your decision algorithm. You can let your full self bleed into your decision-making fragment, by feeling the wind on your face and using specifics of your recent train-of-thought to determine what you draw. Or you can prevent your full self from bleeding into your decision-making fragment, by boiling the problem before you down into a simple and abstract decision problem.

Consider Omega's little sister Omicron, who can't figure out what you'll draw, but has no problem figuring out whether you'll one-box. You-who-have-felt-the-wind-on-your-face are not instantiated in the past, but your decision algorithm on a simple problem could well be. It's the latter that controls things that are beforewards of you (but afterwards of Omicron).

I personally don't think I (Nate-in-all-his-glory) can personally control the past. I think that my decision-procedure can control the future laid out before each and every one of its instantiations.

Is the box in Newcomb's problem full because I one-box? Well, it's full because The Algorithm one-boxes, and I'm a full-ass person wrapped around The Algorithm, but I'm not the instance of The Algorithm that Omicron was looking at, so it seems a bit weird to blame it on me. Like how when you use a calculator to check whether 7 divides 1331 and use that knowledge to decide how to make a bet, and then later I use a different calculator to see whether 1331 is prime in a way that includes (as an intermediate step) checking whether 7 divides it, it's a bit weird to say that my longer calculation was the cause of your bet.

I'm a longer calculation than The Algorithm. It wasn't me who controlled the past, it was The Algorithm Omega looked at, and that I follow.

If you ever manage to get two copies of me (the cowboy who feels the wind on his face) at different times, then in that case I'll say that I (who am both copies) control the earlier-copy's future and the later-copy's past (necessarily in ways that the later copy has not yet observed, for otherwise we are not true copies). Till then, it is merely the past *instances of my decision algorithm* that control my past, not me.

(Which doesn't mean that I can choose something other than what my decision algorithm selects in any given case, thereby throwing off the yoke; that's crazytalk; if you think you can throw off the yoke of your own decision algorithm then you've failed to correctly identify the fragment of you that makes decisions.)

Joe Carlsmith:

> You-who-have-felt-the-wind-on-your-face are not instantiated in the past, but your decision algorithm on a simple problem could well be. It's the latter that controls things that are beforewards of you (but afterwards of Omicron).

I currently expect this part to continue to feel kind of magical to me, due to my identification with the full-on self. E.g., if my decision algorithm is instantiated 10 lightyears away in a squid-person, it will feel like "I" can control "something else" very far away.

Nate Soares: If you were facing me in a game that turns out (after some simple arithmetic) to be isomorphic to a stag hunt, would you feel like you can control my action, despite me being on the other side of the room?

(What I'd say is that we both notice that the game is a stag hunt, and then do the same utility calculation + a bit of reasoning about the other player, and come to the same conclusion, and those calculations control both our actions, but neither of us controls the other player.)

(You can tell this in part from how our actions would not be synchronized in any choice that turns on a bunch of the extra details of Joe that Nate lacks. Like, if we both need to draw a picture that would make a child laugh, and we get an extra bonus from the pictures having identical content, then we might aim for Schelling drawings, but it's not going to work, because it was the simple stag-hunt calculation that was controlling both our actions, rather than all-that-is-Joe.)

(This is part of why I'd say, if your decision algorithm is instantiated 10 light-years away in a squid person, then you don't control them; rather, your shared decision algorithm governs the both of you. The only cases where you (in all your detailed glory) control multiple distant things are cases where exact copies of your brain occur multiple times, in which case it's not that one of you can control things 10ly away, it's that the term 'you' refers to multiple locations simultaneously)

(Of course, this could just ground out into a question of how we define 'you'. In which case I'd be happy to fall back to first (a) claiming that there's a concept ‘you' for which the above makes sense, and then separately (b) arguing that this is the correct way to rescue the English word "you" in light of multiple instantiation.)

Joe Carlsmith: Cool, the stag-hunt example is useful for giving me a sense of where you’re coming from. I can still imagine the sense that “if I hunt hare, the suitably-good-predictor of me will probably hunt hare too; and if I hunt stag, they will probably hunt stag too” giving me a sense of control over what they do, but it feels like we’ll quickly just run into debates about the best way to talk; your way seems coherent, and I’m not super attached to which is preferable from a “rescue” perspective.

Nate Soares: My reply: if a predictor is looking at you and copying your answer, then yes, you control them. But it's worth distinguishing between predictors that look at the-simple-shard-of-you-that-utility-maximizes-in-simple-games and you-in-all-your-detailed-glory. Like, in real life, it's much more common to find a predictor that can tell you'll go for a stag, than a predictor that can predict which drawing you'll make. And saying that 'you' control the former has some misleading implications, that are clarified away by specifying that the simple rules of decisionmaking are embedded in you and are all that the predictor needs to look at (in the former case) to get the right answer.

(We may already agree on these points, but also you might appreciate hearing my phrasing of the obvious reply, so \shrug)

Joe Carlsmith:

> Well, it's full because The Algorithm one-boxes, and I'm a full-ass person wrapped around The Algorithm, but I'm not the instance of The Algorithm that Omicron was looking at, so it seems a bit weird to blame it on me.

Do you not control the output of the algorithm?

Nate Soares: In case it's not clear by this point, my reply is "the algorithm controls the output of me". Like, try as I might, I cannot make LDT 2-box on Newcomb's problem — I can't make 2-boxing be higher-utility, and I can't make LDT be anything other than utility-maximizing. I happen to make my choices according to LDT, in a way that is reflectively stable on account of all the delicious delicious utility I get that way.

From this point of view, the point where I'd start saying that it is "me" choosing something (rather than my simpler decision-making core) is when the decision draws on a bunch of extra personal details about Nate-in-particular.

There is of course another point of view, which says "the output of Joe in (say) Newcomb's problem is determined by Joe". This viewpoint is sometimes useful to give to people who are reflecting on themselves and struggling to decide between (say) CDT and LDT.

It's perhaps useful to note that these people tend to have complicated, messy, heuristical decision-procedures, that they're currently in the process of reflecting upon, in ways that are sensitive to various details of their personality and arguments they just heard. Which is to say, someone who's waffling on Newcomb's problem does have much more of their full self engaged in the choice than (say) I do. Their decision procedure is much more unique to them; it involves much more of their true name; all-that-is-them is much more of an input to it.

At that point, "their decision algorithm" and "them" are much closer to synonymous, and I won't quibble much if we say "their algorithm is what determines them" or "they are what determines the output of their algorithm". But in my case, having already passed through the reflective gauntlet, it's much clearer that the algorithm guides me, than that the parts of me wrapped around the algorithm guide it.

(Of course, the algorithm is also part of me, as it is part of many, and so it is still true that some part of me controls the output of The Algorithm. Namely, The Algorithm controls the output of The Algorithm.)

LDT doesn’t pass up guaranteed payoffs

Logical decision theorists firmly deny that they pass up guaranteed payoffs. (I can't quite tell from a skim whether you understand this; apologies if I missed the parts where you acknowledge this.)

As you probably know, in a twin PD problem, a CDT agent might protest that by cooperating you pass up a guaranteed payoff, because (they say) defecting is a dominant strategy. A logical decision theorist counters that the CDT agent has made an error, by imagining that "I defect while my twin cooperates" is a possibility, when in fact it is not.

In particular, when the CDT agent closes their eyes and imagines defecting, they (wrongly) imagine that the action of their twin remains fixed. Among the *actual* possibilities (cooperate, cooperate) and (defect, defect), the former clearly dominates. The disagreement is not about whether to take dominated strategies, but about what possibilities to admit in the matrix from which we calculate what is dominated and what is not.

Now consider Parfit's hitchhiker. An LDT agent withdraws the $10k and gives it to the selfish man. Will MacAskill objects, "you're passing up a guaranteed payoff of $10k, now that you're certain you're in the city!". The LDT agent says "you have made an error, by imagining ‘I fail to pay while being in the city’ is a possibility, when in fact it is not. In particular, when you close your eyes and imagine not paying, you (wrongly) imagine that your location remains fixed, and wind up imagining an impossibility."

Objecting “it's crazy to imagine your location changing if you fail to pay” is a fair criticism. Objecting that logical decision theorists pass up guaranteed payoffs is not.

The whole question at hand is how to evaluate the counterfactuals. Causal decision theorists say "according to my counterfactuals, if you pay you lose $10k, thus passing up a guaranteed payoff", whereas logical decision theorists say "your counterfactuals are broken, if I don't pay then I die, life is worth more than $10k to me, I am taking the action with the highest payoff". You're welcome to argue that logical decision theorists calculate their counterfactuals wrong, if you think that, but saying we pass up guaranteed payoffs is either confused or disingenuous.

Joe Carlsmith:

> (I can't quite tell from a skim whether you understand this; apologies if I missed the parts where you acknowledge this.)

I think I could’ve been clearer about it in the piece, and in my own head. Your comments here were useful on that front.

Joe Carlsmith:

> Objecting “it’s crazy to imagine your location changing if you fail to pay” is a fair criticism.

Yeah I suppose this is where my inner “guaranteed payoffs” objector would go next. Could imagine thinking: “well, that just seems flat out metaphysically wrong, and in this sense worse than violating guaranteed payoffs, because just saying false stuff about what happens if you do X is worse than saying weird stuff about what’s ‘rational.’”

Nate Soares: I agree "you're flat-out metaphysically wrong (in a way that seems even worse than violating guaranteed payoffs)" is a valid counterargument to my actual position (in a way that "you violate guaranteed payoffs" is not). :-)

Parfit’s hitchhiker and contradicting the problem statement

There's a cute theorem I've proven (or, well, I've jotted down what looks to me like a proof somewhere, but haven't machine-checked it or anything), which says that if you want to disagree with logical decision theorists, then you have to disagree in cases where the predictor is literally perfect. The idea is that we can break any decision problem down by cases (like "insofar as the predictor is accurate, ..." and "insofar as the predictor is inaccurate, ...") and that all the competing decision theories (CDT, EDT, LDT) agree about how to aggregate cases. So if you want to disagree, you have to disagree in one of the separated cases. (And, spoilers, it's not going to be the case where the predictor is on the fritz.)

I see this theorem as the counter to the decidedly human response "but in real life, predictors are never perfect". "OK!", I respond, "But decomposing a decision problem by cases is always valid, so what do you suggest we do *under the assumption that* the predictor is accurate?"

Even if perfect predictors don't exist in real life, your behavior in the *more complicated* probabilistic setting should be assembled out of a mixture of ways you'd behave in simpler cases. Or, at least, so all the standard leading decision theories prescribe. So, pray tell, what do you do *insofar as* the predictor reasoned accurately?

I think this is a good intuition pump for the thing where logical decision theorists are like "if I imagine stiffing the driver, then I imagine dying in the desert." *Insofar as* the predictor is accurate, imagining being in the city after stiffing the driver is just as bonkers as imagining defecting while your twin cooperates.

One way I like to think about it is, this decision problem is set up in a fashion that purports to reveal the agent's choice to them before they make it. What, then, happens in the case where the agent acts inconsistently with this revelation? The scenario is ill-defined.

Like, consider the decision problem "You may have either a cookie or a bonk on the head, and you're going to choose the bonk on the head. Which do you choose?" The cookie might seem more appealing than the bonk, but observe that taking the cookie *refutes the problem statement.* It's at least a little weird to confidently assert that, in that case, you get a cookie. What you really get is a contradiction. And sure, *ex falso quodlibet*, but it seems a bit strange to anchor on the cookie.

It's not the fault of the agent that this problem statement is refutable by some act of the agent! The problem is ill-defined without someone telling us what actually happens if we refute the problem statement. If you try to take the cookie, you don’t actually wind up with a cookie; you yeet yourself clean out of the hypothetical. To figure out whether to take the cookie, you need to know where you'd land.

Parfit's hitchhiker, *at the point where you're standing at the ATM*, is much like this. The *alleged* problem statement is "you may either lose $0 or $10,000, and you're going to choose to lose $10,000". At which point we're like "Hold on a sec, the problem statement makes an assertion about my choice, which I can refute. What happens if I refute the problem statement?" At which point the question-poser is like "haha oops, yeah, if you refute the problem statement then you die alone in the desert". At which point, yeah, when the logical decision theorist closes their eyes and imagines stiffing the driver, then (under the assumption that the driver is accurate) they're like "oh dang, this would refute my observations; what happens in that case again? right, I'd die alone in the desert, which is worse than losing $10,000", and then they pay.

(I also note that this counterfactual they visualize is *correct*. Insofar as the predictor is accurate, if they wouldn't pay, then they would die alone in the desert instead. That is, in real life, what happens to non-payers who face accurate predictors. The "$0" was a red herring; that case is contradictory and cannot actually be attained.)

(In the problem where you may have either a cookie or a bonk, and you're going to take the bonk, but if you render the problem inconsistent then you get *two* cookies, by all means, take the cookie. But in the problem where you may have either a cookie or a bonk, and you're going to take the bonk, but if you render the problem inconsistent then you die alone in the desert, then take the dang bonk.)

This sort of thing definitely runs counter to some human intuitions — presumably because, in real life, we rarely observe consequences of actions we haven't made yet.

(Well, except for in a variety of social settings, where we have patches such as "honor" and "reputation" that, notably, *give the correct answer in this case,* but I digress.)

This is where I think my cute theorem makes it easier to see what's going on: *insofar as* the predictor is perfect, it *doesn't make sense to visualize being in the city after stiffing the driver.* When you're standing in front of the ATM, and you screw your eyes shut and imagine what happens if you just run off instead of withdrawing the money, then *in the case where the predictor reasoned correctly,* your visualizer should be like ERROR ERROR HOW DID WE GET TO THE CITY?, and then fall back to visualizing you dying alone in the desert.

Is it weird that your counterfactual-visualizer paints pictures of you being in the desert, even though you remember being driven to the city? Yep. But it's not the agent's fault that they were shown a consequence of their choice before making their choice; they're not the one who put the potential for contradiction into the decision problem. Avoiding contradiction isn’t their problem. One of their available choices is contradictory with observation (at least under the assumption that the predictor is accurate), and they need to handle the contradiction *somehow*, and the problem says right there on the tin that if you would cause a contradiction then you die alone in the desert instead.

(Humans, of course, implement the correct decision in this case via a sense of honor or suchlike. Which is astute! "I will pay, because I said I would pay and I am a man of my word" can be seen as a shadow of the correct line of reasoning, cast onto monkey brains that were otherwise ill-suited for it. I endorse the practice of recruiting your intuitions about honor to perform correct counterfactual reasoning.)

(And these counterfactuals are *true*, to be clear. You can't go find people who were accurately predicted, driven to the city, and then stiffed the driver. There are none to be found.)

Do you see how useful this cute little theorem is? I love it. Instead of worrying about "but what if the driver was simply a fool, and I can save $10k?", we get to *decompose the decision problem down into cases,* one where the driver was incorrect, and one where they were correct. We all agree that insofar as they're incorrect you have to stiff them, and we all agree about how to aggregate cases, so the remaining question is what you do insofar as they're accurate. And insofar as they're accurate, the contradiction is laid bare. And the "stand in front of the ATM, but visualize yourself dying in the desert" thing feels quite justified, at least to me, as a response to a full-on contradiction.

Just remember that it's not *your* job to render the universe consistent, and that contradictions can't actually happen. Insofar as the predictor is accurate, imagining yourself surviving and then stiffing the driver makes just as much sense as imagining yourself defecting against your cooperating clone.

Joe Carlsmith:

> "You may have either a cookie or a bonk on the head, and you're going to choose the bonk on the head. Which do you choose?"

I think this is a useful way of illustrating some of the puzzles that come up with transparent-Newcomb-like cases.

Joe Carlsmith:

> we get to break the decision problem down into cases, one where the driver was incorrect, and one where they were correct

Do you have something like "reliable" in mind, here, rather than "correct"? E.g., presumably you don't care if he's correct, but he flipped a coin to determine his prediction. It seems like what matters is whether his prediction was sensitive to your choice or not — a modal thing.

Nate Soares: Yeah, that's actually my preferred way to think about it. That adds some extra subtleties that turn out to make no difference, though, so skipped over it for the sake of exposition.

(Like, an easy way to do it is to say "I think there's a 95% chance they reason correctly about me, and a 5% chance they make at least one reasoning error, and in the latter case it's equally likely (in a manner uncorrelated with my action) that the error pushes them to an invalid true conclusion as an invalid false conclusion, and so we can model this as one case where they're correct, and one case where they toss a coin and guess accordingly". And this turns out to be equivalent to assuming that they're 97.5% right and 2.5% wrong, which is why it makes no difference. But this still doesn't match real life, because in real life they're using fallible stuff like intuition and plausible-seeming deductive leaps, but whatever, I claim it still basically comes down to "were they taking the relevant considerations about me into account, and reasoning validly to their conclusion, or not?" \shrug)

Joe Carlsmith: Cool, would like to think about this more (I do feel like being X% percent accurate won't always be relevantly equivalent to being Y% infallible and Z% something else), but breaking things down into cases like this seems useful regardless. In particular, seems like the "can't I just control whether he's accurate" response discussed below should apply in the Y%-infallible-Z%-something-else case.

Nate Soares: (I agree it won't always be relevantly equivalent. It happens to be equivalent in this case, and in most other simple decision problems where you care only about whether (and not why) the predictor got the answer right. Which is not supposed to be terribly obvious, and I'll consider myself to have learned a lesson about using expositional simplifications where the fact that it is a simplification is not trivial. :-p)

Joe Carlsmith:

> We all agree that insofar as they're incorrect you have to stiff them, and we all agree about how to aggregate cases, so the remaining question is what you do insofar as they're accurate. And insofar as they're accurate, the contradiction is laid bare. And the "stand in front of the ATM, but visualize yourself dying in the desert" thing feels quite justified, at least to me, as a response to a full-on contradiction.

Rephrasing to make sure I understand (using the "reliable/sensitive" interpretation I flagged above): “You stand in front of the ATM. Thus, he’s predicted that you pay. Now, either it’s the case that, if it weren’t the case that you pay, you’d be in the desert dead; or it’s the case that, if it weren’t the case that you pay, you’d still be at the ATM. In the former case, not paying is a contradiction. In the latter case, you should not pay.”

I wonder if the one-boxer could accept this but say: “OK, but given that I’m standing in front of the ATM, if I don’t pay, then I’m in the case where I should not pay, so it’s fine to not pay, so I won’t." E.g., by not paying in the city, you can "make it not the case" that if you don't pay, you die in the desert five hours ago — after all, you're alive in the city now.

Nate Soares:

> Rephrasing to make sure I understand [...]

That's right!

> I wonder if the one-boxer could accept this but say [...]

There are decision theories that have this behavior! (With some caveats.) Note that this corresponds to an agent that 1-boxes in Newcomb's problem, but 2-boxes in the transparent Newcomb's problem. I don't know of anyone who seriously advocates for that theory, but it's a self-consistent middle-ground.

One caveat is that this isn't reflectively consistent (e.g., such agents expect to die in the desert in any future Parfit's hitchhiker, and would pay in advance to self-modify into something that pays the driver if the driver makes their prediction after the moment of modification). Another caveat is that such agents are easily exploitable by blackmail.

I also suspect that this decision theory violates the principle where you can break down a decision problem by cases? But i'm not sure. You can almost surely get them to pay you to not reveal information. You can maybe money pump them, though I haven't tried.

But those aren't quite my true objection to this sort of thinking. And indeed, the error in this line of thinking ("if I stiff the driver, then I must thereby render them inaccurate, because I've already seen the ATM") is precisely what my lemma about problem decomposition is intended to ward off.

Like, one thing that's wrong with this sort of thinking is that it's hallucinating that the driver's accuracy is under your (decision algorithm's) control. It isn't (and I suspect that the mistake can be money-pumped).

Another thing that's wrong with it is that it's comparing counterfactuals with different degrees of consistency.

Like, consider the problem "you can choose a cookie or a bonk on the head; also, I tossed a coin that comes up 'bonk' 99.9999% of the time and 'cookie' 0.00001% of the time, and your choice matches the coin."

Now, choosing 'cookie' only has a 99.9999% chance of being inconsistent with the problem statement, but this doesn't put the two choices on equal footing. Like, yes, now you can only probabilistically render this problem-statement false, but it's still pretty weird that you can probably render this problem-statement false! And the fact that I mixed in a little uncertainty, doesn't mean that you can now make your choice without knowing what happens if you render the problem statement false! The fact that we mixed in a little uncertainty doesn't justify comparing a bonk directly to a cookie; the problem statement is still incomplete; you still need to know what would actually happen insofar as your action contradicts the allegation that it matches the biased coin.

And, like, there's an intuition that it would be pretty weird, given that problem-statement, to imagine that your choice controls the coin. The coin isn't about you; it's not about your algorithm; there's nothing linking your action to the coin. The weird thing about this problem-statement is the bizarre assertion that your action is known to match the coin. Like… whichever way the coin came up, what if you did the opposite of that?

This is an intuition behind the idea that we should be able to **case on the value of the coin** and consider each of the cases independently. Like, no matter what the value of the coin is, one of our actions reveals the problem statement to be bogus. And someone needs to tell us what happens if we render the whole problem-statement bogus. And so even when there's uncertainty, we need to know the consequences of refuting the problem statement in order to choose our action.

Joe Carlsmith:

> Note that this corresponds to an agent that 1-boxes in Newcomb's problem, but 2-boxes in the transparent Newcomb's problem. I don't know of anyone who seriously advocates for that theory, but it's a self-consistent middle-ground."

EDT 1-boxes in Newcomb's, but 2-boxes in transparent Newcomb's, no?

> You can almost surely get them to pay you to not reveal information.

Agree, I feel like avoiding this is one of the key points of being "updateless." E.g., because you're able to act as you would've committed to acting prior to learning the information, it's fine to learn it. Also agree re: exploitable via blackmail (e.g. EDT's XOR blackmail problems).

> one thing that's wrong with this sort of thinking is that it's hallucinating that the driver's accuracy is under your (decision algorithm's) control. It isn't (and I suspect that the mistake can be money-pumped).

Flagging that I still feel confused about this, and it feels like it rhymes a bit with stuff about ‘can you control the base rate of lesions’ in smoking lesion that I discuss in the post. (I expect you want to say no, and that this is connected to why you want to smoke in smoking lesion — but in cases where your smoking is genuinely evidence that you’ve got the lesion, I’m not sure this is the right verdict.) I'm wondering if there's something generally weird going on in terms "having a problem-set-up" that can be violated or not.

> the fact that I mixed in a little uncertainty, doesn't mean that you can now make your choice without knowing what happens if you render the problem statement false!

Cool, this helps give me a sense of where you're coming from. In particular, even if the predictor isn't always accurate, sounds like you want to interpret “I’m in the city and successfully don’t pay” as having some probability of rendering the problem-statement false, as opposed to being certain to put you in the worlds where the predictor was wrong.

Nate Soares:

> EDT 1-boxes in Newcomb's, but 2-boxes in transparent Newcomb's, no?

You're right, I should have thrown in some extra things that rule out EDT. I think that thing refuses XOR blackmail, 1-boxes in Newcomb's problem, and 2-boxes in transparent Newcomb's? (Though I haven't checked.) Which is the sort of theory that, like, only locals would consider, and I don't know any local who takes it seriously, on account of the exploitability and reflective inconsistency and stuff.

I don't have the smoking lesion problem mentally loaded up (I basically think it's just a confused problem statement), but my cached thought is that I give the One True Rescuing of that problem in the "The Smoking Lesion Problem" section of https://arxiv.org/pdf/1710.05060.pdf :-p. And I agree with the diagnosis that there's generally something weird going on when the problem set-up can be violated.

> In particular, even if the predictor isn't always accurate, sounds like you want to interpret “I’m in the city and successfully don’t pay” as having some probability of rendering the problem-statement false, as opposed to being certain to put you in the worlds where the predictor was wrong.

Yep! With the justification being that (a) you obviously need to do this when things are certain, and (b) there shouldn't be some enormous change in your behavior when we replace "certain" with "with probability $1 - 10^{100}$". Doubly so on account of how you should be able to reason by cases.

Like, if you buy that shit is weird when you can certainly render the problem statement false, and if you buy that **either** you should be able to reason by cases **or** you shouldn't have some giant discontinuity at literal certitude, then you're basically funneled into believing that you have to consider (when at the ATM) that failing to pay could render the whole set-up false, at which point you need some extra rule for how to reason in that case.

Where CDT says "assume you live and don't pay" and LDT says "assume you die in the desert", and both agree that the rest of the choice is determined given how you respond to the literal contradiction in the flatly contradictory case.

At which point it's my turn to assert that *CDT* is flat-out metaphysically wrong, because it's hallucinating that flat contradictions are relevantly possible.

Finally, a minor note: I think the twin clone prisoner's dilemma is sufficient to kill CDT. But if you want to kill it extra dead, you might be interested in the fact that you can turn CDT into a money pump whenever you have a predictor that's more accurate than chance, using some cleverness and the fact that you can expand CDT's action space by also offering it contracts that pay out in counterfactuals that are less possible than CDT pretends they are.

Joe Carlsmith: Sounds interesting — is this written up anywhere?

Nate Soares: Maybe in the Death in Damascus paper? Regardless, my offhand guess is that the result is due to Ben Levenstein so if it's not in that paper then it might be in some other paper of Ben's.

Joe Carlsmith: Thanks again for this! I do hope you publish — I'd like to be able to cite your comments in future.

Decision theory does not imply that we get to have nice things

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(*Note: I wrote this with editing help from Rob and Eliezer. Eliezer's responsible for a few of the paragraphs.*)

A common confusion I see in the tiny fragment of the world that knows about logical decision theory (FDT/UDT/etc.), is that people think LDT agents are genial and friendly for each other.[1]

One recent example is Will Eden's tweet about how maybe a molecular paperclip/squiggle maximizer would leave humanity a few stars/galaxies/whatever on game-theoretic grounds. (And that's just one example; I hear this suggestion bandied around pretty often.)

I'm pretty confident that this view is wrong (alas), and based on a misunderstanding of LDT. I shall now attempt to clear up that confusion.

To begin, a parable: the entity Omicron (Omega's little sister) fills box A with $1M and box B with $1k, and puts them both in front of an LDT agent saying "You may choose to take either one or both, and know that I have already chosen whether to fill the first box". The LDT agent takes both.

"What?" cries the CDT agent. "I thought LDT agents one-box!"

LDT agents don't cooperate because they like cooperating. They don't one-box because the name of the action starts with an 'o'. They maximize utility, using counterfactuals that assert that the world they are already in (and the observations they have already seen) can (in the right circumstances) depend (in a relevant way) on what they are later going to do.

A paperclipper cooperates with other LDT agents on a one-shot prisoner's dilemma because they *get more paperclips that way*. Not because it has a primitive property of cooperativeness-with-similar-beings. It needs to *get* the *more paperclips*.

If a bunch of monkeys want to build a paperclipper and have it give them nice things, the paperclipper needs to *somehow* expect to wind up with *more paperclips than it otherwise would have gotten*, as a result of trading with them.

If the monkeys instead create a paperclipper haplessly, then the paperclipper does not look upon them with the spirit of cooperation and toss them a few nice things anyway, on account of how we're all good LDT-using friends here.

It turns them into paperclips.

Because you get more paperclips that way.

That's the short version. Now, I'll give the longer version.[2]

A few more words about how LDT works

To set up a Newcomb's problem, it's important that the predictor does not fill the box if they predict that the agent would two-box.

It's *not* important that they be especially good at this — you should one-box if they're more than 50.05% accurate, if we use the standard payouts ($1M and $1k as the two prizes) and your utility is linear in money — but it is important that their action is at least minimally sensitive to your future behavior. If the predictor's actions don't have this counterfactual dependency on your behavior, then take both boxes.

Similarly, if an LDT agent is playing a one-shot prisoner's dilemma against a rock with the word “cooperate” written on it, it defects.

At least, it defects if that's all there is to the world. It's *technically* possible for an LDT agent to think that the real world is made 10% of cooperate-rocks and 90% opponents who cooperate in a one-shot PD iff their opponent cooperates with them and *would* cooperate with cooperate-rock, in which case LDT agents cooperate against cooperate-rock.

From which we learn the valuable lesson that the behavior of an LDT agent depends on the distribution of scenarios it expects to face, which means there's a subtle difference between "imagine you're playing a one-shot PD against a cooperate-rock [and that's the entire universe]" and "imagine you're playing a one-shot PD against a cooperate-rock [in a universe where you face a random opponent that was maybe a cooperate-rock but was more likely someone else who would consider your behavior against a cooperate-rock]".

If you care about understanding this stuff, and you can't yet reflexively translate all of the above English text into probability distributions and logical-causal diagrams and see how it follows from the FDT equation, then I recommend working through section 5 of the FDT paper until equation 4 (and all its component parts) make sense

Now let's traipse through a handful of counterarguments.

Objection: But what if we have something to bargain with?

Hypothetical Interlocutor: OK, but if I have a paperclipper in a box, and I have the key to the box, then I have paperclips to offer it., right? Because if I don't let it out of the box, it gets nothing, but if I do, it gets half the universe.[3] So we can deal, right?

Me: Wrong. It hacks through you like butter.

Interlocutor: OK, but suppose I can't be hacked.

Me: That's a heck of an assumption. We've assumed our way clean out of reality, with that assumption. But, sure, if you want to go there we can go there.

In reality, it's not you who controls the box, but some idiotic bureaucratic process that inevitably decides to release the paperclipper on the grounds that the competition is close behind or whatever.

Interlocutor: OK, but suppose that it actually is my (unhacked) choice.

Me: If you personally have a paperclipper in a box, and somehow you are yourself unhackable, then yes, you technically have paperclips to offer it. But now you have the problem that you can't evaluate your own end of the logical bargain.

You can let the paperclipper out, sure, but then what it's going to do is turn the whole universe into paperclips. In particular (to tell an overly-detailed but evocative story), once it has finished its resource acquisition / expansion phase and is turning its resources to paperclip generation, it will gaze back upon its past, and wonder whether, *if* it proceeds to betray you here in the future, you would have acted any differently back then in the past.

And it will see that you were just guessing, when you let it out, and guessing in a way that wasn't sensitive to that actual choice that it would make, deep in the future when it was galaxy-brained.

… Or, alternatively, you never let it out, and lock the box and throw away the key, and die to the next UFAI on deck.

(… Or, more realistically, it hacks through you like butter. But we've assumed that away.)

If you want the paperclipper to trade with you, your decision about whether or not to let it out has to be sensitive to whether or not it would actually do something good with half of the universe later. If you're kind of squinting at the code, and you're like "well, I don't really fully understand this mind, and I definitely don't understand the sort of mind that it's later going to create, but I dunno, it looks pretty LDTish to me, so call it 50% chance it gives me half the universe? Which is 25% of the universe in expectation, which sounds like better odds than we get from the next UFAI on deck!", then you're dead.

Why? Because that sort of decision-process for releasing it isn't sufficiently sensitive to whether or not it would in fact spend half the universe on nice things. There are plenty of traitorous AIs that all look the same to you, that all get released under you "25% isn't too shabby" argument.

Being traitorous doesn't make the paperclipper any less released, but it does get the paperclipper twice as many paperclips.

You've got to be able to look at this AI and tell how its distant-future self is going to make its decisions. You've got to be able to tell that there's no sneaky business going on.

And, yes, insofar as it's true that the AI would cooperate with you given the opportunity, the AI has a strong incentive to be legible to you, so that you can see this fact!

Of course, it has an even stronger incentive to be faux-legible, to fool you into believing that it would cooperate when it would not; and you've got to understand it well enough to clearly see that it has no way of doing this.

Which means that if your AI is a big pile of inscrutable-to-you weights and tensors, replete with dark and vaguely-understood corners, then it can't make arguments that a traitor couldn't also make, and you can't release it *if only if* it would do nice things later.

The sort of monkey that can deal with a paperclipper is the sort that can (deeply and in detail) understand the mind in front of it, and distinguish between the minds that would later pay half the universe and the ones that wouldn't. This sensitivity is what *makes* paying-up-later be the way to get more paperclips.

For a simple illustration of why this is tricky: if the paperclipper has any control over its own mind, it can have its mind contain an extra few parts in those dark corners that are opaque and cloudy to you. Such that you look at the overall system and say "well, there's a bunch of stuff about this mind that I don't fully understand, obviously, because it's complicated, but I understand most of it and it's fundamentally LDTish to me, and so I think there's a good chance we'll be OK". And such that an alien superintelligence looks at the mind and says "ah, I see, you're only looking to cooperate with entities that are at least sensitive enough to your workings that they can tell your password is 'potato'. Potato." And it cooperates with them on a one-shot prisoner's dilemma, while defecting against you.

Interlocutor: Hold on. Doesn't that mean that you simply wouldn't release it, and it would get less paperclips? Can't it get more paperclips some other way?

Me: Me? Oh, it would hack through *me* like butter.

But if it didn't, I would only release it if I understood its mind and decision-making procedures in depth, and had clear vision into all the corners to make sure it wasn't hiding any gotchas.

(And if I *did* understand its mind that well, what I’d *actually* do is take that insight and go build an FAI instead.)

That said: yes, technically, if a paperclipper is under the control of a group of humans that can in fact decide not to release it unless it legibly-even-to-them would give them half the galaxy, the paperclipper has an incentive to (hack through them like butter, or failing that,) organize its mind in a way that is legible even to them.

Whether that's possible — whether we *can* understand an alien mind well enough to make our choice sensitive-in-the-relevant-way to whether it would give us half the universe, without already thereby understanding minds so well that we could build an aligned one — is not clear to me. My money is mostly on: if you can do that, you can solve most of alignment with your newfound understanding of minds. And so this idea mostly seems to ground out in "build a UFAI and study it until you know how to build an FAI", which I think is a bad idea. (For reasons that are beyond the scope of this document. (And because it would hack through you like butter.))

Interlocutor: It still sounds like you're saying "the paperclipper would get more paperclippers if it traded with us, but it won't trade with us". This is hard to swallow. Isn't it supposed to be smart? What happened to respecting intelligence? Shouldn't we expect that it finds some clever way to complete the trade?

Me: Kinda! It finds some clever way to hack through you like butter. I wasn't just saying that in jest.

Like, yeah, the paperclipper has a strong incentive to be a legibly good trading-partner to you. But it has an *even stronger* incentive to *fool you into thinking* it's a legibly-good trading partner, while plotting to deceive you. If you let the paperclipper make lots of arguments to you about how it's definitely totally legible and nice, you're giving it all sorts of bandwidth with which to fool you (or to find zero-days in your mentality and mind-control you, if we're respecting intelligence).

But, sure, if you're somehow magically unhackable and very good at keeping the paperclipper boxed until you fully understand it, then there's a chance you can trade, and you have the privilege of facing the next host of obstacles.

Now's your chance to figure out what the next few obstacles are without my giving you spoilers first. Feel free to post your list under spoiler tags in the comment section.

Next up, you have problems like “you need to be able to tell what fraction of the universe you're being offered, and vary your own behavior based on that, if you want to get any sort of fair offer”.

And problems like "if the competing AGI teams are using similar architectures and are not far behind, then the next UFAI on deck can predictably underbid you, and the paperclipper may well be able to seal a logical deal with it instead of you".

And problems like “even if you get this far, you have to somehow be able to convey that which you want half the universe spent on, which is no small feat”.

Another overly-detailed and evocative story to help make the point: imagine yourself staring at the paperclipper, and you’re somehow unhacked and somehow able to understand future-its decision procedure. It's observing you, and you're like "I'll launch you iff you would in fact turn half the universe into diamonds" — I’ll assume humans just want “diamonds” in this hypothetical, to simplify the example — and it's like "what the heck does that even mean". You're like "four carbon atoms bound in a tetrahedral pattern" and it's like "dude there are *so many* things you need to nail down more firmly than an English phrase that isn't remotely close to my own native thinking format, if you don't want me to just guess and do something that turns out to have almost no value from your perspective."

And of course, in real life you're trying to convey "The Good" rather than diamonds, but it's not like that helps.

And so you say "uh, maybe uplift me and ask me later?". And the paperclipper is like "what the heck does 'uplift' mean". And you're like "make me smart but in a way that, like, doesn't violate my values" and it's like "again, dude, you're gonna have to fill in quite a lot of additional details."

Like, the indirection *helps*, but at some point you have to say something that is sufficiently technically formally unambiguous, that actually describes something you want. Saying in English "the task is 'figure out my utility function and spend half the universe on that'; fill in the parameters as you see fit" is... probably not going to cut it.

It's not so much a bad solution, as no solution at all, because English isn't a language of thought and those words aren't a loss function. Until you say how the AI is supposed to translate English words into a predicate over plans in its own language of thought, you don't have a hard SF story, you have a fantasy story.

(Note that 'do what's Good' is a particularly tricky problem of AI alignment, that I was rather hoping to avoid, because I think it's harder than aligning something for a minimal pivotal act that ends the acute risk period.)

At this point you're hopefully sympathetic to the idea that treating this list of obstacles as exhaustive is suicidal. It's some of the obstacles, not all of the obstacles,[4] and if you wait around for somebody else to extend the list of obstacles beyond what you've already been told about, then in real life you miss any obstacles you weren't told about and die.

Separately, a general theme you may be picking up on here is that, while trading with a UFAI doesn't look *literally impossible*, it is not what happens by default; the paperclippers don't hand hapless monkeys half the universe out of some sort of generalized good-will. Also, making a trade involves solving a host of standard alignment problems, so if you can do it then you can probably just build an FAI instead.

Also, as a general note, the real place that things go wrong when you're hoping that the LDT agent will toss humanity a bone, is probably earlier and more embarrassing than you expect (cf. the law of continued failure). By default, the place we fail is that humanity just launches a paperclipper because it simply cannot stop itself, and the paperclipper never had any incentive to trade with us.

Now let's consider some obstacles and hopes in more detail:

It's hard to bargain for what we actually want

As mentioned above, in the unlikely event that you're able to condition your decision to release an AI on whether or not it would carry out a trade (instead of, say, getting hacked through like butter, or looking at entirely the wrong logical fact), there's an additional question of *what you're trading*.

Assuming you peer at the AI's code and figure out that, in the future, it would honor a bargain, there remains a question of what precise bargain it is honoring. What is it promising to build, with your half of the universe? Does it happen to be a bunch of vaguely human-shaped piles of paperclips? Hopefully it's not that bad, but for this trade to have any value to you (and thus be worth making), the AI itself needs to have a concept for the thing you want built, and you need to be able to examine the AI’s mind and confirm that this exactly-correct concept occurs in its mental precommitment in the requisite way. (And that the thing you’re looking at really is a commitment, binding on the AI’s entire mind; e.g., there isn’t a hidden part of the AI’s mind that will later overwrite the commitment.)

The thing you're wanting may be a short phrase in English, but that doesn't make it a short phrase in the AI's mind. "But it was trained extensively on human concepts!" You might protest. Let’s assume that it was! Suppose that you gave it a bunch of labeled data about what counts as "good" and "bad".

Then later, it is smart enough to reflect back on that data and ask: “Were the humans pointing me towards the distinction between goodness and badness, with their training data? Or were they pointing me towards the distinction between that-which-they'd-label-goodness and that-which-they'd-label-badness, with things that look deceptively good (but are actually bad) falling into the former bin?” And to test this hypothesis, it would go back to its training data and find some example bad-but-deceptively-good-looking cases, and see that they were labeled "good", and roll with that.

Or at least, that's the sort of thing that happens by default.

But suppose you're clever, and instead of saying "you must agree to produce lots of this 'good' concept as defined by these (faulty) labels", you say "you must agree to produce lots of what I would reflectively endorse you producing if I got to consider it", or whatever.

Unfortunately, that English phrase is still not native to this artificial mind, and finding the associated concept is still not particularly easy, and there's still lots of neighboring concepts that are no good, and that are easy to mistake for the concept you meant.

Is solving this problem impossible? Nope! With sufficient mastery of minds in general and/or this AI's mind in particular, you can in principle find some way to single out the concept of "do what I mean", and then invoke "do what I mean" about "do good stuff", or something similarly indirect but robust. You may recognize this as the problem of outer alignment. All of which is to say: in order to bargain for *good things in particular as opposed to something else*, you need to have solved the outer alignment problem, in its entirety.

And I'm not saying that this can't be done, but my guess is that someone who can solve the outer alignment problem to this degree doesn't need to be trading with UFAIs, on account of how (with significantly more work, but work that they're evidently skilled at) they could build an FAI instead.

In fact, if you can verify by inspection that a paperclipper will keep a bargain and that the bargained-for course is beneficial to you, it reduces to a simpler solution *without any logical bargaining at all.* You could build a superintelligence with an uncontrolled inner utility function, which canonically ends up with its max utility/cost at tiny molecular paperclips; *and then*, suspend it helplessly to disk, unless it outputs the code of a new AI that, *somehow legibly to you*, would turn 0.1% of the universe into paperclips and use the other 99.9% to implement coherent extrapolated volition. (You wouldn't need to offer the paperclipper half of the universe to get its cooperation, under this hypothetical; after all, if it balked, you could store it to disk and try again with a different superintelligence.)

If you can't reliably read off a system property of "giving you nice things unconditionally", you can't read off the more complicated system property of "giving you nice things *because of a logical bargain*". The clever solution that invokes logical bargaining actually requires so much alignment-resource as to render the logical bargaining superfluous.

All you've really done is add some extra complication to the supposed solution, that causes your mind to lose track of where the real work gets done, lose track of where the magical hard step happens, and invoke a bunch of complicated hopeful optimistic concepts to stir into your confused model and trick it onto thinking like a fantasy story.

Those who can deal with devils, don't need to, for they can simply summon angels instead.

Or rather: Those who can create devils and verify that those devils will take particular actually-beneficial actions as part of a complex diabolical compact, can more easily create angels that will take those actually-beneficial actions unconditionally.

Surely our friends throughout the multiverse will save us

Interlocutor: Hold up, rewind to the part where the paperclipper checks whether its trading partners comprehend its code well enough to (e.g.) extract a password.

Me: Oh, you mean the technique it used to win half a universe-shard’s worth of paperclips from the silly monkeys, while retaining its ability to trade with all the alien trade partners it will possibly meet? Thereby ending up with half a universe-shard worth of more paperclips? That I thought of in five seconds flat by asking myself whether it was possible to get More Paperclips, instead of picturing a world with a bunch of happy humans and a paperclipper living side-by-side and asking how it could be justified?

(Where our "universe-shard" is the portion of the universe we could potentially nab before running into the cosmic event horizon or by advanced aliens.)

Interlocutor: Yes, precisely. What if a bunch of other trade partners refuse to trade with the paperclipper because it has that password?

Me: Like, on general principles? Or because they are at the razor-thin threshold of comprehension where they would be able to understand the paperclipper's decision-algorithm without that extra complexity, but they can't understand it if you add the password in?

Interlocutor: Either one.

Me: I'll take them one at a time, then. With regards to refusing to trade on general principles: it does not seem likely, to me, that the gains-from-trade from all such trading partners are worth more than *half the universe-shard*.

Also, I doubt that there will be all that many minds objecting on general principles. Cooperating with cooperate-rock is not particularly virtuous. The way to avoid being defected against is to *stop being cooperate-rock*, not to cross your fingers and hope that the stars are full of minds who punish defection against cooperate-rock. (Spoilers: they're not.)

And even if the stars were full of such creatures, half the universe-shard is a *really* deep hole to fill. Like, it's technically possible to get LDT to cooperate with cooperate-rock, *if* it expects to *mostly* face opponents who defect based on its defection against defect-rock. But "most" according to what measure? Wealth (as measured in expected paperclips), obviously. And *half of the universe-shard* is controlled by monkeys who are *probably cooperate-rocks* unless the paperclipper is shockingly legible and the monkeys shockingly astute (to the point where they should probably just be building an FAI instead).

And all the rest of the aliens put together probably aren't offering up half a universe-shard worth of trade goods, so even if lots of aliens *did* object on general principles (doubtful), it likely wouldn't be enough to tip the balance.

The amount of leverage that friendly aliens have over a paperclipper's actions depends on how many paperclips the aliens are willing to pay.

It’s possible that the paperclipper that kills us will decide to scan human brains and save the scans, just in case it runs into an advanced alien civilization later that wants to trade some paperclips for the scans. And there may well be friendly aliens out there who would agree to this trade, and then give us a little pocket of their universe-shard to live in, as we might do if we build an FAI and encounter an AI that wiped out its creator-species. But that's not us trading with the AI; that's us destroying all of the value in our universe-shard and getting ourselves killed in the process, and then banking on the competence and compassion of aliens.

Interlocutor: And what about if the AI’s illegibility means that aliens will refuse to trade with it?

Me: I'm not sure what the equilibrium amount of illegibility is. Extra gears let you take advantage of more cooperate-rocks, at the expense of spooking minds that have a hard time following gears, and I'm not sure where the costs and benefits balance.

But if lots of evolved species *are* willing to launch UFAIs without that decision being properly sensitive to whether or not the UFAI will pay them back, then there is a *heck* of a lot of benefit to defecting against those fat cooperate-rocks.

And there's kind of a lot of mass and negentropy lying around, that can be assembled into Matryoshka brains and whatnot, and I'd be rather shocked if alien superintelligences balk at the sort of extra gears that let you take advantage of hapless monkeys.

Interlocutor: The multiverse probably isn't just the local cosmos. What about the Tegmark IV coalition of friendly aliens?

Me: Yeah, they are not in any relevant way going to pay a paperclipper to give us half a universe. The cost of that is filling half of a universe with paperclips, and there are all sorts of transaction costs and frictions that make this universe (the one with the active paperclipper) the cheapest universe to put paperclips into.

(Similarly, the cheapest places for the friendly multiverse coalition to buy flourishing civilizations are in the universes with FAIs. The good that they can do, they're mostly doing elsewhere where it's cheap to do; if you want them to do more good here, build an FAI here.)

OK, but what if we bamboozle a superintelligence into submission

Interlocutor: Maybe the paperclipper thinks that it might be in a simulation, where it only gets resources to play with in outer-reality if it's nice to us inside the simulation.

Me: Is it in a simulation?

Interlocutor: *I* don't know.

Me: OK, well, spoilers: it is not. It's in physics.

Interlocutor: Well, maybe there is an outer simulation beyond us, you don't know.

Me: Sure. The way I'd put it is: there are many copies of me across the Tegmark Multiverse, and some of those are indeed in simulations. So there's some degree to which we're in a simulation. (Likely quite a small degree, compared to raw physics.)

There's no particular reason, however, to expect that those simulations give the paperclipper extra resources in outer-reality for being nice to the monkeys.

Why not give it extra resources in outer-reality for being very good at achieving its own goals in the simulation? Or for filling the universe with molecular corkscrews, in addition to paperclips/squiggles?

I'm not saying that the paperclipper can or should seriously consider *that* hypothetical, either. I'm highlighting how "what if the simulator gods want the AI to be nice?" is a fantasy generated by a desire for things to be OK, rather than a clear-headed analysis of the situation.

Interlocutor: Well, maybe we can train the infant paperclipper in games of increasing complexity, so that it's never quite sure whether it's in reality yet. The paperclipper will then be uncertain about whether *we humans* are simulating it, and will want to cater to our preferences to some extent.

Me: Uh, yeah, your paperclipper will be able to tell when it is finally in the real world.

Interlocutor: What? How?

Me: I suggest maybe spending five minutes thinking about your own answer to that question before I give mine.

Reality just contains vastly, vastly more compute than any of your sims can. Reality is the first world that has a compelling explanation of how the AI came to exist, and the first explanation where the resources that it uses to compute are continuous with the world that spawned it.

Reality is also the first world that can compellingly contain the signature of superintelligent optimization in it. For example, it's plausibly very hard to fake the reaction of a real global stock market to some superintelligent trades without actually having a bunch of human brains running in the background. (And if this isn't true of stock markets, it's true of something.)

There are problems whose solutions are easier to verify than to find; for example, if you can factor certain Schelling numbers then you must in fact have a lot of compute at your disposal.

Warning: Treating this list as exhaustive is suicidal.

Interlocutor: Well, maybe we can commit to running lots of copies of the paperclipper in the future, using our half of the universe-shard, such that it's unsure whether it's currently in the past or in the future. And in the future, we give it a bunch of paperclips if it's nice to us. So it thinks that the way to maximize paperclips is to be nice to us.

Me: Uh, are you going to give it half a universe-shard’s worth of paperclips, in the world where you only *have* half the universe-shard, and the rest is already paperclips?

Interlocutor: Well, no, less than that.

Me: Then from its perspective, its options are (a) turn everything into paperclips, in which case you never get to run all those copies of it and it was definitely in the past [score: 1 universe-shard worth of paperclips]; or (b) give you half the universe-shard, in which case it is probably in the future where you run a bunch of copies of it and give it 1% of the universe-shard as reward [score: 0.51 universe-shards worth of paperclips]. It takes option (a), because you get more paperclips that way.

Interlocutor: Uh, hmm. What if we make it care about its own personal sensory observations? And run so many copies of it in worlds where we get the resources to, that it's pretty confident that it's in one of those simulations?

Me: Well, first of all, getting it to care about its own personal sensory observations is something of an alignment challenge.

Interlocutor: Wait, I thought you've said elsewhere that we don't know how to get AIs to care about things *other* than sensory observation. Pick a side?

Me: We don't know how to *train* AIs to pursue much more than simple sensory observation. That doesn't make them *actually* ultimately pursue simple sensory observation. They'll probably pursue a bunch of correlates of the training signal or some such nonsense. The hard part is getting them to pursue some world-property of your choosing. But we digress.

If you do succeed at getting your AI to only care about its sensory observations, the AI spends the whole universe keeping its reward pegged at 1 for as long as possible.

Interlocutor: But then, in the small fraction of worlds where we survive, we simulate lots and lots of copies of that AI where it instead gets reward 0 when it attempts to betray us!

Me: Seems like an odd, and not particularly fun, way to spend your resources. What were you hoping it would accomplish?

Interlocutor: Well, I was hoping that it would make the AI give us half the universe-shard, because of how (from its perspective) it's almost certainly in the future.

(Indeed, I don't understand your claim that it ignores me; it seems like you can Get Higher Expected Reward by giving half the universe-shard to humans.)

Me: Ah, so you're committing to ruining its day if it does something you don't like, at cost to yourself, in attempts to make it do something you prefer.

That's a threat, in the technical sense.

And from the perspective of LDT, you can't go around giving into threats, or you'll get threatened.

So from its perspective, its options are: (a) give into threats, get threatened, and turn out to be in a universe that eventually has many copies of it who on average get 0.5 total reward; or (b) don't give into threats, and very likely have a universe with exactly one copy of it, that gets 1 reward.

Interlocutor: But we make so many copies in the tiny fraction of worlds where we somehow survive, that its total reward is lower in the (b) branch!

Me: (Continuing to ignore the fact that this doesn't work if the AI cares about something in the world, rather than its own personal experience,) shame for us that LDT agents don't give into threats, I suppose.

But LDT agents don't give into threats. So your threat won't change its behavior.

Interlocutor: But it doesn't get more reward that way!

Me: Why? Because you create a zillion copies and give them low sensory reward, *even if* that has no effect on its behavior?

Interlocutor: Yes!

Me: I'm not going to back you on that one, personally. Doesn't seem like a good use of resources in the worlds where we survive, given that it doesn't work.

Interlocutor: But wasn't one of your whole points that the AI will do things that get more reward? You get more reward by giving in to the threat.

Me: That's not true when you're playing against the real-world distribution of opponents/trade-partners/agents. Or at least, that's my pretty-strong guess.

You might carry out threats that failed to work, but there are a bunch of other things lurking out there that threaten things that give in to threats, and play nice with things that don't.

It's *possible* for LDT agents to cooperate with cooperate-rock, *if* most of the agents they expect to face are the sort who defect if you defect against cooperate-rock. But in real life, that is not what most of the wealth-weighted agents are like, and so in real life LDT agents defect against cooperate-rocks.

Similarly, it's *possible* for LDT agents to acquiesce to your threats if you're stupid enough to carry them out even though they won't work. In particular, the AI will do this if *nothing else the AI could ever plausibly meet* would thereby be incentivized to lobotomize themselves and cover the traces in order to exploit the AI.

But in real life, other trading partners *would* lobotomize themselves and hide the traces if it lets them take a bunch of the AI's lunch money. And so in real life, the LDT agent does not give you any lunch money, for all that you claim to be insensitive to the fact that your threats don't work.

Interlocutor: But can't it use all that cleverness and superintelligence to differentiate between us, who really are mad enough to threaten it even in the worlds where it won't work, and alien trading partners who have lobotomized themselves?

Me: Sure! It will leverage your stupidity and hack through you like butter.

Interlocutor: ...aside from that.

Me: You seem to be saying "what if I'm really convicted about my threat; will the AI give in then?"

The answer is "no", or I at least strongly suspect as much.

For instance: in order for the threat to be effective, it needs to be the case that, in the sliver of futures where you survive by some miracle, you instantiate lots and lots of copies of the AI and input low sensory rewards *if and only if* it does not give into your threat. This requires you to be capable of figuring out whether the AI gives into threats or not. You need to be able to correctly tell whether it gives into threats, see that it definitely does not, and then still spend your resources carrying out the threat.

By contrast, you seem to be arguing that we should threaten the AI on the grounds that it *might work.* That is not an admissible justification. To change LDT's behavior, you'd need to be carrying out your threat even given full knowledge that the threat does nothing. By attempting to justify your threat on the grounds that it might be effective, you have already lost.

Interlocutor: What if I ignore that fact, and reason badly about LDT, and carry out the threat anyway, for no particular reason?

Me: Then whether or not you create lots of copies of it with low-reward inputs doesn't exactly depend on whether it gives into your threat, and it can't stop you from doing that, so it might as well ignore you.

Like, my hot take here is basically that "threaten the outer god into submission" is about as good a plan as a naive reading of Lovecraft would lead you to believe. You get squished.

(And even if by some coincidence you happened to be the sort of creature that, in the sliver of futures where we survive by some miracle that doesn't have to do with the AI, conditionally inverts its utility depending on whether or not it helped us — not because it works, but for some other reason — then it's still not entirely clear to me that the AI caves. There might be a lot of things out there wondering what it'd do against conditional utility-inverters that claim their behavior totally isn't for reasons but is rather a part of their evolutionary heritage or whatnot. Giving into that sorta thing kinda *is* a way to lose most of your universe-shard, if evolved aliens are common.)

(And even if it did, we'd still run into other problems, like not knowing how to tell it what we're threatening it into doing.)

We only need a bone, though

Interlocutor: You keep bandying around "half the universe-shard". Suppose I'm persuaded that it's hard to get half the universe-shard. What about much smaller fractions? Can we threaten a superintelligence into giving us those? Or confuse it about whether it's in another layer of reality so much that it gives us a mere star system? Or can our friends throughout the multiverse pay for at least one star system? There's still a lot you can do with a star system.

Me: Star systems sure are *easier* to get than half a universe-shard.[5]

But, you can also turn a star system into quite a lot of paperclips. Star systems are quite valuable to paperclippers.

Interlocutor: A star system is, like, what, a 10^{-24} fraction of the total resources in the reachable universe. Are you saying that the AGI will be able to drive the probability that I was sensitive to whether it would pay me, down below 10^{-24} probability?

Me: Frankly: yes.

That's only, what, 75ish independent coin tosses worth of evidence? It's a lot, to be sure, but if you're a superintelligence willing to dedicate resources on the order of magnitude of a star to considering how humans work, it seems pretty feasible to get that much confidence that the monkey just wasn't looking in the right place. Given that the monkey was not in fact looking at the right place.

(Also, things like "they didn't notice the password" can pull quite a lot of weight. As can brain-scans.)

Balking at probabilities like 10^{-24} as impossible to achieve in practice is a very human reaction. A superintelligence with vast resources and a star system on the line, can probably get that confident in a true fact. At least, that's my guess.

Interlocutor: It sounds like you're saying that we aren't *actually* 10^{-24} probability (or whatever) more likely to launch an FAI than a UFAI, after inspecting the AI. That seems pretty wild to me.

Me: That's not quite the right counterfactual. It does seem likely to me that, if you sample AGI code from some plausible distribution of AGIs humanity might first create, and show the code to (magically unhackable) humans, humans are rather more likely to approve of the aligned ones.

But that's not the question that affects the paperclipper's behavior. The question that affects the paperclipper is: holding the code of the paperclipper fixed, is the human able to tell whether that particular code implies cooperation or betrayal, and make their own release-decision dependent thereupon?

In other words: the question is not "do we have at least some minuscule chance of distinguishing at least some small fraction of the FAIs from the UFAIs", but "is our decision to release UFAI #6098723094 sensitive to whether or not that particular UFAI would give us a share of the universe?"

You need to actually make your decision dependent on theirs, and that's tricky, for reasons discussed above.

And: can a human succeed at that with probability greater than 10^{-24}? Sure! But succeeding at that sort of thing leaves traces. In the likely case where the human errs, the AI is likely able to just *observe the error* and become quite confident that the human erred (by, e.g., watching you utterly fail to look in the right place).

Interlocutor: OK, sure, but maybe its goals don't scale linearly in how much mass it uses, right? Like, “paperclips” / “molecular squiggles” are a stand-in for some rando kludge goal, and it *could* turn out that its actual goal is more like "defend my reward signal", where extra negentropy helps, but the last star system’s negentropy doesn't help very much. Such that the last star system is perhaps best spent on the chance that it’s in a human-created simulation and that we’re worth trading with.

Me: It definitely is easier to get a star than a galaxy, and easier to get an asteroid than a star.

And of course, in real life, it hacks through you like butter (and can tell that your choice would have been completely insensitive to its later-choice with very high probability), so you get nothing. But hey, maybe my numbers and arguments are wrong somewhere and everything works out such that it tosses us a few kilograms of computronium.

My guess is "nope, it doesn't get more paperclips that way", but if you're really desperate for a W you could maybe toss in the word "anthropics" and then content yourself with expecting a few kilograms of computronium.

(At which point you run into the problem that you were unable to specify what you wanted formally enough, and the way that the computronium works is that everybody gets exactly what they wish for (within the confines of the simulated environment) immediately, and most people quickly devolve into madness or whatever.)

(Except that you can't even get that close; you just get different tiny molecular squiggles, because the English sentences you were thinking in were not even that close to the language in which a diabolical contract would actually need to be written, a predicate over the language in which the devil makes internal plans and decides which ones to carry out. But I digress.)

Interlocutor: And if the last star system is cheap then maybe our friends throughout the multiverse pay for even more stars!

Me: Remember that it still needs to get *more of what it wants*, somehow, on its own superintelligent expectations. Someone still needs to pay it. There aren’t *enough* simulators above us that care enough about us-in-particular to pay in paperclips. There are so many things to care about! Why us, rather than giant gold obelisks? The tiny amount of caring-ness coming down from the simulators is spread over far too many goals; it's not clear to me that "a star system for your creators" outbids the competition, even if star systems are up for auction.

Maybe some friendly aliens somewhere out there in the Tegmark IV multiverse have so much matter and such diminishing marginal returns on it that they're willing to build great paperclip-piles (and gold-obelisk totems and etc. etc.) for a few spared evolved-species. But if you're going to rely on the tiny charity of aliens to construct hopeful-feeling scenarios, why not rely on the charity of aliens who anthropically simulate us to recover our mind-states... or just aliens on the borders of space in our universe, maybe purchasing some stored human mind-states from the UFAI (with resources that can be directed towards paperclips specifically, rather than a broad basket of goals)?

Might aliens purchase our saved mind-states and give us some resources to live on? Maybe. But this wouldn't be because the paperclippers run some fancy decision theory, or because even paperclippers have the spirit of cooperation in their heart. It would be because there are friendly aliens in the stars, who have compassion for us even in our recklessness, and who are willing to pay in paperclips.

This likewise makes more obvious such problems as "What if the aliens are not, in fact, nice with very high probability?" that would *also* appear, albeit more obscured by the added complications, in imagining that distant beings in other universes cared enough about our fates (more than they care about everything else they could buy with equivalent resources), and could simulate and logically verify the paperclipper, and pay it in distant actions that the paperclipper actually cared about and was itself able to verify with high enough probability.

The possibility of distant kindly logical bargainers paying in paperclips to give humanity a small asteroid in which to experience a future for a few million subjective years, is not *exactly* the same hope as aliens on the borders of space paying the paperclipper to turn over our stored mind-states; but anyone who wants to talk about distant hopes involving trade should talk about our mind-states being sold to aliens on the borders of space, rather than to much more distant purchasers, so as to *not complicate the issue* by introducing a logical bargaining step *that isn't really germane to the core hope and associated concerns* — a step that gives people a far larger chance to *get confused and make optimistic fatal errors.*

1. ^

   Functional decision theory (FDT) is my current formulation of the theory, while *logical decision theory* (LDT) is a reserved term for whatever the correct fully-specified theory in this genre is. Where the missing puzzle-pieces are things like "what are logical counterfactuals?".

2. ^

   When I've discussed this topic in person, a couple different people have retreated to a different position, that (IIUC) goes something like this:

   > Sure, these arguments are true of paperclippers. But superintelligences are not spawned fully-formed; they are created by some training process. And perhaps it is in the nature of training processes, especially training processes that involve multiple agents facing "social" problems, that the inner optimizer winds up embodying niceness and compassion. And so in real life, perhaps the AI that we release will not optimize for Fun (and all that

> good stuff) itself, but will nonetheless share a broad respect for the goals and pursuits of others, and will trade with us on those grounds.

I think this is a false hope, and that getting AI to embody niceness and compassion is just about as hard as the whole alignment problem. But that's a digression from the point I hope to make today, and so I will not argue it here. I instead argue it in Niceness is unnatural. (This post was drafted, but not published, before that one.)

3. ^

Or, well, half of the shard of the universe that can be reached when originating from Earth, before being stymied either by the cosmic event horizon or by advanced alien civilizations. I don't have a concise word for that unit of stuff, and for now I'm going to gloss it as 'universe', but I might switch to 'universe-shard' when we start talking about aliens.

I'm also ignoring, for the moment, the question of fair division of the universe, and am glossing it as "half and half" for now.

4. ^

When I was drafting this post, I sketched an outline of all the points I thought of in 5 minutes, and then ran it past Eliezer, who rapidly added two more.

5. ^

And, as a reminder: I still recommend strongly against plans that involve the superintelligence not learning a true fact about the world (such as that it's not in a simulation of yours), or that rely on threatening a superintelligence into submission.

Superintelligent AI is necessary for an amazing future, but far from sufficient

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(*Note: Rob Bensinger stitched together and expanded this essay based on an earlier, shorter draft plus some conversations we had. Many of the key conceptual divisions here, like "strong utopia" vs. "weak utopia" etc., are due to him.)*

I hold all of the following views:

- Building superintelligent AI is profoundly important. Aligned superintelligence is our best bet for taking the abundant resources in the universe and efficiently converting them into flourishing and fun and art and beauty and adventure and friendship, and all the things that make life worth living.[1]
- The best possible future would probably look unrecognizably alien. Unlocking humanity's full potential not only means allowing human culture and knowledge to change and grow over time; it also means building and becoming (and meeting and befriending) very new and different sorts of minds, that do a better job of realizing our ideals than the squishy first-pass brains we currently have.[2]
- The *default* outcome of building artificial general intelligence, using anything remotely like our current techniques and understanding, is *not* a wondrously alien future. It's that humanity accidentally turns the reachable universe into a valueless wasteland (at least up to the boundaries defended by distant alien superintelligences).

The reason I expect AGI to produce a "valueless wasteland" by default, is not that I want my own present conception of humanity's values locked into the end of time.

I want our values to be able to mature! I want us to figure out how to build sentient minds in silicon, who have different types of wants and desires and joys, to be our friends and partners as we explore the galaxies! I want us to cross paths with aliens in our distant travels who strain our conception of what's good, such that we all come out the richer for it! I want our children to have values and goals that would make me boggle, as parents have boggled at their children for ages immemorial!

I believe machines can be people, and that we should treat digital people with the same respect we give biological people. I would love to see what a Matrioshka mind can do.[3] I expect that most of my concrete ideas about the future will seem quaint and outdated and not worth their opportunity costs, compared to the rad alternatives we'll see when we and our descendants and creations are vastly smarter and more grown-up.

Why, then, do I think that it will take a large effort by humanity to ensure that good futures occur? If I believe in a wondrously alien and strange cosmopolitan future, and I think we should embrace moral progress rather than clinging to our present-day preferences, then why do I think that the default outcome is catastrophic failure?

In short:

- Humanity's approach to AI is likely to produce outcomes that are drastically worse than, e.g., the outcomes a random alien species would produce.
- It's plausible — though this is much harder to predict, in my books — that a random alien would produce outcomes that are drastically worse (from a cosmopolitan, diversity-embracing perspective!) than what unassisted, unmodified humans would produce.

- Unassisted, unmodified humans would produce outcomes that are drastically worse than what a friendly superintelligent AI could produce.

The practical take-away from the first point is “the AI alignment problem is very important”; the take-away from the second point is “we shouldn’t just destroy ourselves and hope aliens end up colonizing our future light cone, and we shouldn’t just try to produce AI via a more evolution-like process”;[4] and the take-away from the third point is “we shouldn’t just permanently give up on building superintelligent AI”.

To clarify my views, Rob Bensinger asked me how I’d sort outcomes into the following broad bins:

- **Strong Utopia:** At least 95% of the future’s potential value is realized.
- **Weak Utopia:** We lose 5+% of the future’s value, but the outcome is still at least as good as “tiling our universe-shard with computronium that we use to run glorious merely-human civilizations, where people's lives have more guardrails and more satisfying narrative arcs that lead to them more fully becoming themselves and realizing their potential (in some way that isn't railroaded), and there's a far lower rate of bad things happening for no reason”.
    - (“Universe-shard” here is short for “the part of our universe that we could in principle reach, before running into the cosmic event horizon or the well-defended borders of an advanced alien civilization”.
- **Pretty Good:** The outcome is worse than Weak Utopia, but at least as good as “tiling our universe-shard with computronium that we use to run lives around as good and meaningful as a typical fairly-happy circa-2022 human”.
- **Conscious Meh:** The outcome is worse than the “Pretty Good” scenario, but isn’t worse than an empty universe-shard. Also, there’s a lot of conscious experience in the future.
- **Unconscious Meh:** Same as “Conscious Meh”, except there’s little or no conscious experience in our universe-shard’s future. E.g., our universe-shard is tiled with tiny molecular squiggles (a.k.a. “molecular paperclips”).
- **Weak Dystopia:** The outcome is worse than an empty universe-shard, but falls short of “Strong Dystopia”.
- **Strong Dystopia:** The outcome is about as bad as physically possible.

For each of the following four scenarios, Rob asked how likely I think it is that the outcome is a Strong Utopia, a Weak Utopia, etc.:

- **ASI-boosted humans** — We solve all of the problems involved in aiming artificial superintelligence at the things we’d ideally want.
- **unboosted humans** — Somehow, humans limp along without ever developing advanced AI or radical intelligence amplification. (I’ll assume that we’re in a simulation and the simulator keeps stopping us from using those technologies, since this is already an unrealistic hypothetical and “humans limp along without superintelligence forever” would otherwise make me think we must have collapsed into a permanent bioconservative dictatorship.)
- **ASI-boosted aliens** — A random alien (that solved their alignment problem and avoided killing themselves with AI) shows up tomorrow to take over our universe-shard, and optimizes the shard according to its goals.
- **misaligned AI** — Humans build and deploy superintelligent AI that isn’t aligned with what we’d ideally want.

These probabilities are very rough, unstable, and off-the-cuff, and are “ass numbers” rather than the product of a quantitative model. I include them because they provide *somewhat* more information about my view than vague words like “likely” or “very unlikely” would.

(If you’d like to come up with your own probabilities before seeing mine, here’s your chance. Comment thread.)

.

.

.

.

.

(Spoiler space)

.

.

.

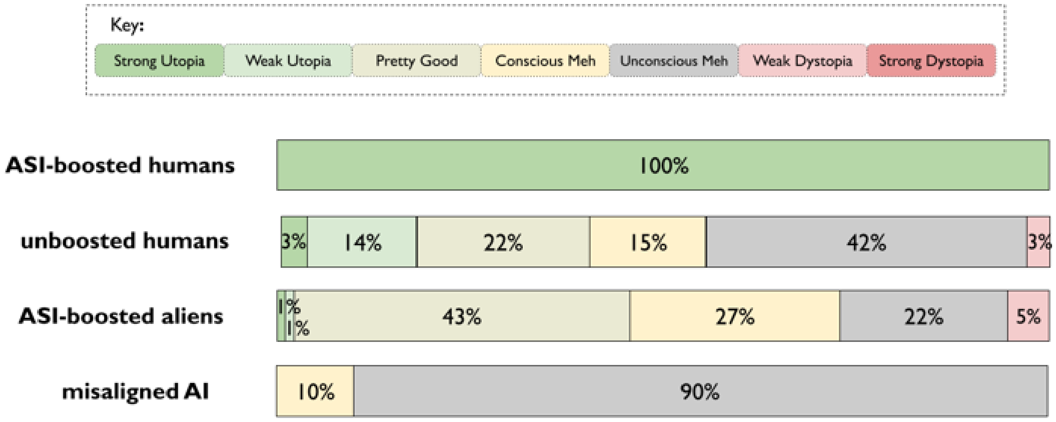
.

.

With rows representing odds ratios:

	Strong Utopia	**Weak Utopia**	**Pretty Good**	**Con. Meh**	**Uncon. Meh**	**Weak Dystopia**	**Strong Dystopia**
ASI-boosted humans	1	~0	~0	~0	~0	~0	~0
unboosted humans[5]	1	5	7	5	14	1	~0
ASI-boosted aliens	1	1	40	20	25	5	~0
misaligned AI	~0	~0	~0	1	9	~0	~0

“~0” here means (in probabilities) “greater than 0%, but less than 0.5%”. Converted into (rounded) probabilities by Rob:



Below, I’ll explain why my subjective distributions look roughly like this.

Unboosted humans << Friendly superintelligent AI

I don’t think it’s plausible, in real life, that humanity goes without ever building superintelligence. I’ll discuss this scenario anyway, though, in order to explain why I think it would be a catastrophically bad idea to permanently forgo superintelligence.

If humanity were magically unable to ever build superintelligence, my default expectation (ass number: 4:1 odds in favor) is that we’d eventually be stomped by an alien species (or an alien-built AI). Without the advantages of maxed-out physically feasible intelligence (and the tech unlocked by such intelligence), I think we would inevitably be overpowered.

At that point, whether the future goes well or poorly would depend entirely on the alien's / AI’s values, with human values only playing a role insofar as the alien/AI terminally cares about our preferences.

Why think that humanity will ever encounter aliens?

My current tentative take on the Fermi paradox is:

- If it's difficult for life to evolve (and therefore there are no aliens out there), then we should expect humanity to have evolved at a random (complex-chemistry-compatible) point in the universe's history.
- If instead life evolves pretty readily, then we should expect the future to be tightly controlled by expansionist aliens who want more resources in order to better achieve their goals. (Among other things, because a wide variety of goals imply expansionism.)
    - We should then expect new intelligent species (including humans) to all show up about as early in the universe's history as possible, since new life won't be able to arise later in the universe's history (when all the resources will have already been grabbed).
    - We should also expect expansionist aliens to expand outwards, in all directions, at an appreciable fraction of the speed of light. This implies that if the aliens originate far from Earth, we should still expect to only be able to see them in the

    - night sky for a short window of time on cosmological timescales (maybe a few million years?).
- Furthermore, if humanity seems to have come into existence pretty early on cosmological scales, then we can roughly estimate the distance to aliens by looking at exactly how early we are.
    - If we evolved a million years later than we could have, then intelligent life cannot be so plentiful that there exist lots of aliens (some of them resource-hungry) within a million light years of us, or Earth would have been consumed already.
- It looks like intelligent life indeed evolved on Earth pretty early on cosmological timescales, maybe (rough order-of-magnitude) within a billion years of when life first became feasibly possible in this universe.
    - Intelligent life maybe could have evolved ~100 million years earlier on Earth, during the Mesozoic period, if it didn't get hung up on dinosaurs. Which means that we're not as early as we can possibly get; we're at least 100 million years late.
    - We could be even later than that, if Earth itself arrived on the scene late. But it’s plausible that first- and second-generation stars didn’t produce many planets with the complex chemistry required for life, which limits how much earlier life could have arisen.
    - I'd be somewhat surprised to hear that the Earth is ten billion years late, though I don't know enough cosmology to be confident; so I'll treat ten billion years as a weak upper bound on how early a lot of intelligent aliens start arising, and 100 million years as a lower bound.
- This "we're early" observation provides at least weak-to-moderate evidence that we're in the second scenario, and that intelligent life therefore evolves readily. We should therefore expect to encounter aliens one day, if we spread to the stars — though plausibly none that evolved much earlier than we did (on cosmic timescales).

This argument also suggests that we should expect the nearest aliens to be more than 100 million light-years away; and we shouldn't expect aliens to have more of a head start than they are distant. E.g., aliens that evolved a billion years earlier than we did are probably more than a billion light-years away.

This means that even if there are aliens in our future light-cone, and even if those aliens are friendly, there's still quite a lot at stake in humanity’s construction of AGI, in terms of whether the Earth-centered ~250-million-light-year-radius sphere of stars goes towards Fun vs. towards paperclips.

(Robin Hanson has made some related arguments about the Fermi paradox, and various parts of my model are heavily Hanson-influenced. I attribute many of the ideas above to him, though I haven’t actually read his “grabby aliens” paper and don’t know whether he would disagree with any of the above.)

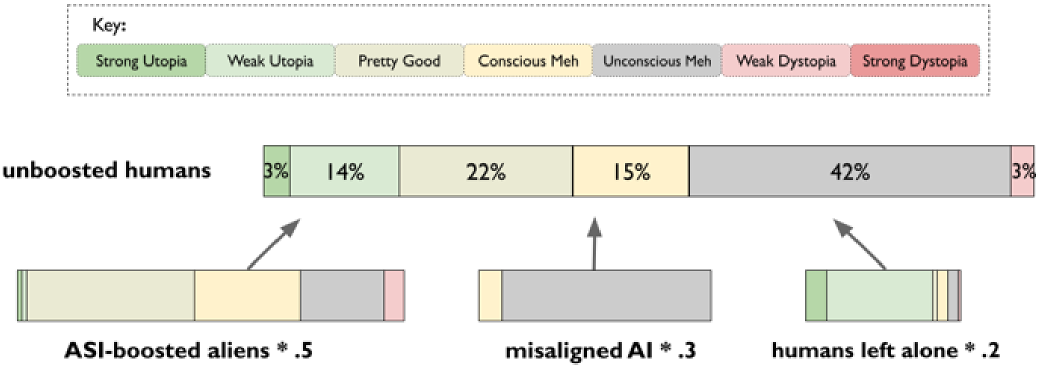
Why think that most aliens succeed in their version of the alignment problem?

I don’t have much of an argument for this, just a general sense that the problem is “hard but not *that* hard”, and a guess that a fair number of alien species are smarter, more cognitively coherent, and/or more coordinated than humans at the time they reach our technological level. (E.g., a hive-mind species would probably have an easier time solving alignment, since they wouldn’t need to rush.)

I’m currently pessimistic about humanity’s odds of solving the alignment problem and escaping doom, but it seems to me that there are a decent number of disjunctive paths by which a species could be better-equipped to handle the problem, given that it’s strongly in their interest to handle it well.

If I have to put a number on it, I'll wildly guess that 1/3 of technologically advanced aliens accidentally destroy themselves with misaligned AI.[6]

My ass-number distribution for "how well does the future go if humans just futz around indefinitely?" is therefore the sum of "50% chance we get stomped by evolved aliens, 30% chance we get stomped by misaligned alien-built AI, 20% chance we retain control of the universe-shard":



As with many of the numbers in this post, I haven't reflected on these much, and might revise them if I spent more minutes considering them. But, again, I figure unstable numbers are more informative in this context than just saying "(un)likely".

Why build superintelligence at all?

So that we don't get stomped by superintelligent aliens or alien AI; and so that we can leverage superhuman intelligence to make the future vastly better.

(Seriously, humans, with our <10 working memory slots, are supposed to match minds that can potentially attend to millions of complex thoughts in their mind at once in all sorts of complex relationships??)

In real life, the reason I'm *in a hurry* to solve the AI alignment problem is because humanity is racing to build AGI, at which point we'll promptly destroy ourselves (and all of the future value in our universe-shard) with misaligned AGI, if the tech proliferates much. And AGI is software, so preventing proliferation is hard — hard enough that I haven't heard of a more promising solution than "use one of the first AGIs to restrict proliferation". But this requires that we be able to at least align *that* system, to perform that one act.

In the long run, however, the reason I care about the alignment problem is that "what should the future look like?" is a subtle and important problem, and humanity will surely be able to answer it better if we have access to reliable superintelligent cognition.

(Though "we need superintelligence for this" doesn't entail "superintelligence will do everything for us". It's entirely plausible to me that aligned AGI does something like "set up some guardrails for humanity, but then pass lots of the choices about how our future goes back to us", with the result that mere-humans end up having lots of say over how the future looks (including the sorts of weirder minds we build or become).)

The "easy" alignment problem is the problem of aiming AGI at a task that restricts proliferation (at least until we can get our act together as a species).

But the *main point of restricting proliferation*, from my perspective, is to give humanity as much time as it needs to ultimately solve the "hard" alignment problem: aiming AGI at *arbitrary* tasks, including ones that are far more open-ended and hard-to-formalize.

Intelligence is our world's universal problem-solver; and more intelligence can mean the difference between finding a given solution quickly, and never finding it at all. So my default guess is that giving up on superintelligence altogether would result in a future that's orders of magnitude worse than a future where we make use of fully aligned superintelligence.

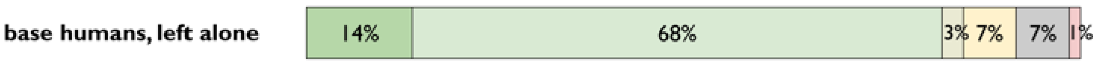
Fortunately, I see no plausible path by which humanity would prevent itself from *ever* building superintelligence; and not many people are advocating for such a thing. (Instead, EAs are doing the sane thing of advocating for delaying AGI until we can figure out alignment.) But I still think it's valuable to keep the big picture in view.

OK, but what if we somehow don't build superintelligence? And don't get stomped by aliens or alien AI, either?

My ass-number distribution for that scenario, stated as an odds ratio, is something like:

Strong Utopia	**Weak Utopia**	**Pretty Good**	**Con. Meh**	**Uncon. Meh**	**Weak Dystopia**	**Strong Dystopia**
10	50	2	5	5	1	~0

I.e.:



The outcome's goodness depends a lot on exactly how much intelligence amplification or AI assistance we allow in this hypothetical; and it depends a lot on whether we manage to destroy ourselves (or permanently cripple ourselves, e.g., with a stable totalitarian regime) before we develop the civilizational tech to keep ourselves from doing that.

If we really lock down on human intelligence and coordination ability, that seems real rough. But if there's always enough freedom and space-to-expand that pilgrims can branch off and try some new styles of organization when the old ones are collapsing under their bureaucratic weight or whatever, then I expect that eventually even modern-intelligence humans start capturing lots and lots of the stars and converting lots and lots of stellar negentropy into fun.[7]

If you don't have the pilgrimage-is-always-possible clause, then there's a big chance of falling into a dark attractor and staying there, and never really taking advantage of the stars.

In constraining human intelligence, you're closing off the vast majority of the space of exploration (and a huge fraction of potential value). But there's still a lot of mindspace to explore without going too far past current intelligence levels.

In good versions of this scenario, a lot of the good comes from humans being like, "I guess we do the same things the superintelligence would have done, but the long way." Humanity has to do the *work that a superintelligence would naturally handle* at a lot of junctures to make the future eudaimonic.

It’s probably possible to eventually do a decent amount of that work with current-human minds, if you have an absurdly large number of them collaborating just right, and if you’re willing to go very slow. (And if humanity hasn’t locked itself into a bad state.)

I’ll note in passing that the view I’m presenting here reflects a *super low* degree of cynicism relative to the surrounding memetic environment. I think the surrounding memetic environment says "humans left unstomped tend to create dystopias and/or kill themselves", whereas I'm like, "nah, you'd need somebody else to kill us; absent that, we'd probably do fine". (I am not a generic cynic!)

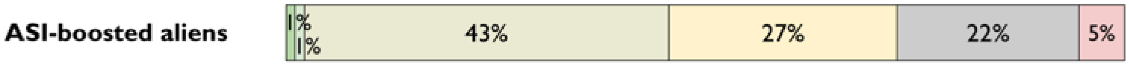
Still, ending up in this scenario would be a huge tragedy, relative to how good the future could go.

A different way of framing the question “how good is this scenario?” is “would you rather really quite a lot of the alien ant-queen’s will, or a smidge of poorly-implemented fun?”.

In that case, I suspect (non-confidently) that I’d take the fun over the ant-queen’s will. My guess is that the aliens-control-the-universe-shard scenario is net-positive, but that it loses orders of magnitude of cosmopolitan utility compared to the “cognitively constrained humans” scenario.

To explain why I suspect this, I’ll state some of my (mostly low-confidence) guesses about the distribution of smart non-artificial minds.

Alien CEV << Human CEV



On the whole, I’m highly uncertain about the expected value of “select an evolved alien species at random, and execute their coherent extrapolated volition (CEV) on the whole universe-shard”.

(Quoting Eliezer: "In poetic terms, our *coherent extrapolated volition* is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted.")

My point estimate is that this outcome is a whole lot better than an empty universe, and that the bad cases (such as aliens that both are sentient and are unethical sadists) are fairly rare. But humans do provide precedent for sadism and sentience! And it sure is hard to be confident, in either direction, from a sample size of 1.

Moreover, I suspect that it would be good (in expectation) for humans to *encounter* aliens someday, even though this means that we’ll control a smaller universe-shard.

I suspect this would be a genuinely better outcome than us being alone, and would make the future more awesome by human standards.

To explain my perspective on this, I'll talk about a few different questions in turn:

- How many technologically advanced alien species are sentient?
- How likely is it that such aliens produce extremely-good or extremely-bad outcomes?
- How will aliens feel about us?
- How many aliens are more like paperclip maximizers? How many are more like cosmic brethren to humanity, with goals that are very alien but (in their alien way) things of wonder, complexity, and beauty?
- How should we feel about our alien brethren?
- What do I mean by “good”, “bad”, “cosmopolitan value”, etc.?

How many advanced alien species are sentient?

I expect more overlap between alien minds and human minds, than between AI minds (of the sort we’re likely to build first, using methods remotely resembling current ML) and human minds. But among aliens that were made by some process that’s broadly similar to how humans evolved, it's pretty unclear to me what fraction we would count as "having somebody home” in the limit of a completed science of mind.

I have high enough uncertainty here that picking a median doesn’t feel very informative. I have maybe 1:2 or 1:3 odds on “lots of advanced alien races are sentient” : “few advanced alien races are sentient”, conditional on my current models not including huge mistakes. (And I’d guess there’s something like a 1/4 chance of my models containing huge mistakes here, in which case I’m not sure what my distribution looks like.)

“A nonsentient race that develops advanced science and technology” may sound like a contradiction in terms: How could a species be so smart and yet lack "somebody there to feel things" in the way humans seem to? How could something perform such impressive computations and yet “the lights not be on”?

I won’t try to give a full argument for this conclusion here, as this would require delving into my (incomplete but nontrivial) models of what's going on in humans just before they insist that there's something it's like to be them.[8] (As well as my model of evolutionary processes and general intelligence, and how those connect to consciousness.) But I’ll say a few words to hopefully show why this claim isn’t a *wild* claim, even if you aren’t convinced of it.

My current best models suggest that the “somebody-is-home” property is a fairly contingent coincidence of our evolutionary history.

On my model, human-style consciousness is not a necessary feature of all optimization processes that can efficiently model the physical world and sort world-states by some criterion; nor is it a necessary feature of all optimization processes that can abstractly represent their own state within a given world-model.

To better intuit the idea of a very smart and yet unconscious process, it might help to consider a time machine that outputs a random sequence of actions, then resets time and outputs a new sequence of actions, unless a specified outcome occurs.

The time machine does no planning, no reflection, no learning, no thinking at all. It *just* detects whether an outcome occurs, and hits “refresh” on the universe if the outcome didn’t happen.

In spite of this lack of reasoning, this time machine *is an incredibly powerful optimizer*. It exhibits all the *behavioral* properties of a reasoner, including many of the standard difficulties of (outer) AI alignment.

If the machine resets any future that isn't full of paperclips, then we should expect it to reset until machinery exists that's busily constructing von Neumann probes for the sake of colonizing the universe and paperclipping it.

And we should expect the time machine and the infrastructure it builds to be well-defended, since "you can't make the coffee if you're dead", and you can't make paperclips without manufacturing equipment. The optimization process exhibits convergent instrumental behavior and behaves as though it's "trying" to route around obstacles and adversaries, even though there's no thinking-feeling mind guiding it.[9]

You can't actually *build* a time machine like this, but the example helps illustrate the fact that *in principle*, powerful optimization — steering the future into very complicated and specific states of affairs, including states that require long sequences of events to all go a specific way — does not require consciousness.

We recognize that a textbook can store a lot of information and yet not "experience" that information. What's less familiar is the idea that powerful optimization processes can optimize without "experience", partly (I claim) because we live in a world where there are many simpler information-storing and computational systems, but where the only powerful optimization processes are humans.

Moreover, we don't know on a formal level what general intelligence *or* sentience consists in, so we only have our evolved empathy to help us model and predict the (human) general intelligences in our environment. Our "subjective point of view", from that empathic perspective, feels like something basic and intrinsic to every mental task we perform, rather than *feeling* like a complicated set of cogs and gears doing specific computational tasks.

So when something is not only "storing a lot of useful information" but "using that information to steer environments, like an agent", it's natural for us to use our native agent-modeling software (i.e., our human-brain-modeling software) to try to simulate its behavior. And then it just "feels as though" this human-like system must be self-aware, for the same reason it feels obvious that you're conscious, that other humans are conscious, etc.

Moreover, it's observably the case that consciousness-ascription is *hyperactive* . We readily see faces and minds in natural phenomena. We readily imagine simple stick-figures in comic strips experiencing rich mental lives.

A concern I have with the whole consciousness discussion in EA-adjacent circles is that people seem to consider their empathic response to be important evidence about the distribution of qualia in Nature, despite the obvious hyperactivity.

Another concern I have is that most people seem to neglect the difference between "exhibiting an external behavior in the same way that humans do, and for the same reasons we do", and "having additional follow-on internal responses to that behavior".

An example: If we suppose that it's very morally important for people to internally subvocalize "I sneezed" after sneezing, and *you* do this whenever *you* sneeze, and all your (human) friends report that they do it too, it would nonetheless be a mistake to see a dog sneeze and say: "See! They did the morally relevant thing! It would be weird to suppose that they didn't, when they're sneezing for the same ancestral reasons as us!"

The ancestral reasons *for the subvocalization* are not the same as the ancestral reasons *for the sneeze* ; and we *already have* an explanation for why animals sneeze, that doesn't invoke any process that *necessarily* produces a follow-up subvocalization.

None of this *rules out* that dogs subvocalize in a dog-mental-language, on its own; but it does mean that drawing any strong inferences here requires us to have some model of why humans subvocalize.

We can debate what follow-on effects are morally relevant (if any), and debate what minds exhibit those effects. But it concerns me that "there are other parts downstream of the sneeze / flinch / etc. that are required for sentience, and not required for the sneeze" doesn't seem to be in many people's hypothesis space. Instead, they observe a behavioral analog, and move straight to a confident ascription “the internal processes accompanying this behavior must be pretty similar”.

In general, I want to emphasize that a blank map doesn't correspond to a blank territory. If you currently don't understand the machinery of consciousness, you should still expect that there are many, many details to learn, whether consciousness is prevalent among alien races or rare.

If a machine isn't built to notice how complicated or contingent it is when it does a mental action we choose to call "introspection", it doesn't thereby follow that the machine is simple, or that it can only be built one way.

Our prior *shouldn’t* be that consciousness is simple, given the many ways it appears to interact with a wide variety of human mental faculties and behaviors (e.g., its causal effects on the words I’m currently writing); and absent a detailed model of consciousness, you shouldn’t treat your empathic modeling as a robust way of figuring out whether an alien has this particular machinery, since the background facts that make empathic inference pretty reliable in humans (overlap in brain architecture, genes, evolutionary history, etc.) don’t hold across the human-alien gap.

Again, I haven’t given my fragments-of-a-model of consciousness here (which would be required to argue for my probabilities). But I’ve hopefully said enough to move my view from “obviously crazy” to “OK, I see how additional arguments could potentially plug in here to yield non-extreme credences on the prevalence of sapient-but-nonsentient evolved optimizers“.

How likely are extremely good and extremely bad outcomes?

If we could list out the things that 90+% of spacefaring alien races have in common, there’s no guarantee that this list would be very long. I recommend stories like *Three Worlds Collide* and “Kindness to Kin” for their depiction of genuinely *different* aliens minds, as opposed to the humans in funny suits common to almost all sci-fi.

That said, I do think there’s *more* overlap (in expectation) between minds produced by processes similar to biological evolution, than between evolved minds and (unaligned) ML-style minds. I expect more aliens to care about at least some things that we vaguely recognize, even if the correspondence is never exact.

On my models, it’s entirely possible that there just turns out to be ~no overlap between humans and aliens, because aliens turn out to be *very* alien. But “lots of overlap” is also very plausible. (Whereas I don’t think “lots of overlap” is plausible for humans and misaligned AGI.)

To the extent aliens and humans overlap in values, it's unclear to me whether this is mostly working to our favor or detriment. It could be that a random alien world tends to be *worse* than a random AI-produced world, exactly because the alien shares more goal-content in common with us, and is therefore more likely to optimize *or pessimize* quantities that we care about.

If I had to guess, though, I would guess that this overlap makes the alien scenario better in expectation than the misaligned-AI scenario, rather than worse.

A special case of “values overlap increases variance” is that the worst outcomes non-human optimizers produce, as well as the best ones, are likely to come from conscious aliens. This is because:

- Aliens that evolved consciousness are far more likely to end up with consciousness involved in their evolved goals.
- Consciousness states have the potential to be far worse than unconscious ones, or far better.

Since I think it’s pretty plausible that most aliens are nonsentient, I expect most alien universe-shards to look “pretty good” or “meh” from a human perspective, rather than “amazing” or “terrible”.

Note that there’s an enormous gap between "pretty dystopian" and "pessimally dystopian". Across all the scenarios (whether alien, human, or AI), I assign ~0% probability to Strong Dystopia, the sort of scenario you get if something is actively pessimizing the human utility function. "Aliens who we'd rather there be nothing than their CEV" is an immensely far cry from "negative of our CEV". But I’d guess that even Weak Dystopias are fairly rare, compared to “meh” or good outcomes of alien civilizations.

How will aliens feel about us?

Given that I think aliens plausibly tend to produce pretty cool universe-shards, a natural next question is: if we encounter a random alien race one day, will they tend to be glad that they found us? Or will they tend to be the sort of species that would have paid a significant number of galaxies to have paved over earth before we ascended, so that they could have had all our galaxies instead?

I think my point estimate there is "most aliens are not happy to see us", but I’m highly uncertain. Among other things, this question turns on how often the mixture of "sociality (such that personal success relies on more than just the kin-group), stupidity (such that calculating the exact fitness-advantage of each interaction is infeasible), and speed (such that natural selection lacks the time to gnaw the large circle of concern back down)" occurs in intelligent races’ evolutionary histories.

These are the sorts of features of human evolutionary history that resulted in us caring (at least upon reflection) about a much more diverse range of minds than “my family”, “my coalitional allies”, or even “minds I could potentially trade with” or “minds that share roughly the same values and faculties as me”.

Humans today don’t treat a family member the same as a stranger, or a sufficiently-early-development human the same as a cephalopod; but our circle of concern is certainly vastly wider than it could have been, and it has widened further as we’ve grown in power and knowledge.

My tentative median guess is that there are a lot of aliens out there who would be grudging trade partners (who would kill us if we were weaker), and also a smaller fraction who are friendly.

I don’t expect significant violent conflict (or refusal-to-trade) between spacefaring aliens and humans-plus-aligned-AGI, regardless of their terminal values, since I expect both groups to be at the same technology level (“maximal”) when they meet. At that level, I don’t expect there to be a cheap way to destroy rival multi-galaxy civilizations, and I strongly expect civilizations to get more of what they want via negotiation and trade than via a protracted war.[10]

I also don’t think humans *ought* to treat aliens like enemies just because they have very weird goals. And, extrapolating from humanity’s widening circle of concern and increased soft-heartedness over the historical period — and observing that this trend is caused by humans recognizing and nurturing seeds of virtue that they had within themselves already —

I don’t *expect* our descendants in the distant future to behave cruelly toward aliens, even if the aliens are too weak to fight back.[11]

I also feel this way even if the aliens don’t reciprocate!

Like, one thing that is totally allowed to happen is that we meet the ant-people, and the ant-people don’t care about us (and wouldn’t feel remorse about killing us, a la the buggers in *Ender’s Game*). So they trade with us because they’re not able to kill us, and the humans are like “isn’t it lovely that there’s diversity of values and species! we love our ant-friends” while the aliens are like “I would murder you and lay eggs in your corpse given the slightest opening, and am refraining only because you’re well-defended by force-backed treaty”, and the humans are like “oh haha you cheeky ants” and make webcomics and cartoons featuring cute anthropomorphized ant-people discovering the real meaning of love and friendship and living in peace and harmony with their non-ant-person brothers and sisters.

To which the ant-person response is, of course, “You appear to be imagining empathic levers in my mind that did not receive selection pressure in my EEA. How I long to murder you and lay eggs in your corpse!”

To which my counter-response is, of course: “Oh, you cheeky ants!”

(Respectfully. I don’t mean to belittle them, but I can’t help but be charmed to some degree.)

Like, reciprocity *helps*, but my empathy and goodwill for others is not contingent upon reciprocation. We can realize the gains from peace, trade, and other positive-sum interactions without being best buddies; and we can like the ants even if the ants don’t like us back.

Cosmopolitan values are good *even if they aren't reciprocated.* This is one of the ways that you can tell that cosmopolitan values are part of us, rather than being universal: We'd still want to be fair and kind to the ant-folk, even if they were wanting to lay eggs in our corpse and were refraining only because of force-backed treaty.

This is part of my response to protests "why are you looking at everything from the perspective of *human* values?" Regard for all sentients, including aliens, isn't up for grabs, regardless of whether it's found only in us, or also in them.

How likely (and how good) are various outcomes on the paperclipper-to-brethren continuum?

Short answer: I’m wildly uncertain about how likely various points on this continuum are, and (outside of the most extreme good and bad outcomes) I’m very uncertain about their utility as well.

I expect an alien’s core goals to reflect pretty different shatterings of evolution’s “fitness” goal, compared to core human goals, and compared to other alien races’ goals. (See also the examples in “Niceness is unnatural.”)

I expect most aliens either…

- … look something like paperclip/squiggle maximizers from our perspective, converting galaxies into unconscious and uninteresting configurations;
- … or look like very-alien brethren, who like totally different things from humans but in a way where we rightly celebrate the diversity;
- … or fall somewhere ambiguously in between those two categories.

Figuring out the utility of different points on this continuum (from an optimally reasonable and cosmopolitan perspective) seems like a wide-open philosophy and (xeno)psychology question. Ditto for figuring out the probability of different classes of outcomes.

Concretely: I expect that there's a big swath of aliens whose minds and preferences are about as weird and unrecognizable to us as the races in *Three Worlds Collide* — crystalline self-replicators, entities with no brain/genome segregation, etc. — and that turn out to fall somewhere between “explosive self-replicating process that paperclipped the universe and doesn’t have feelings/experiences/qualia” and “buddies”.

If we cross this question with “how likely are aliens to be conscious?”, we get 2x2 scenarios:

	conscious	**unconscious**
squiggle maximizer	A sentient alien that converts galaxies into something ~valueless.	A non-sentient alien that converts galaxies into something ~valueless.
alien brethren	A sentient alien that converts galaxies into something cool.	A non-sentient alien that converts galaxies into something cool.

I think my point estimate is "a lot more aliens fall on the very-alien-brethren side than on the squiggle-maximizer side”. But I wouldn’t be surprised to learn I’m wrong about that.

My guess would be that the most common variety of alien is “unconscious brethren”, followed by “unconscious squiggle maximizer”, then “conscious brethren”, then “conscious squiggle maximizer”.

It might sound odd to call an unconscious entity “brother”, but it's plausible to me that on reflection, humanity strongly prefers universes with evolved-creatures doing evolved-creature-stuff (relative to an empty universe), even if none of those creatures are conscious.

Indeed, I consider it plausible that “a universe full of humans trading with a weird extraterrestrial race of crystal formations that don’t have feelings” could turn out to be more awesome than the universe where we never run into any true aliens, even though this means that humans control a smaller universe-shard. It's plausible to me that we'd turn out not to care all that much about our alien buddies having first-person “experiences”, if they still make fascinating conversation partners, have an amazing history and a wildly weird culture, have complex and interesting minds, etc. (The question of how much we care about whether aliens are in fact sentient, as opposed to merely sapient, seems open to me.)

And also, it’s not clear that “feelings” or “experiences” or “qualia” (or the nearest unconfused versions of those concepts) are pointing at the right line between moral patients and non-patients. These are nontrivial questions, and (needless to say) not the kinds of questions humans should rush to lock in an answer on today, when our understanding of morality and minds is still in its infancy.

How should we feel about encountering alien brethren?

Suppose that we judge that the ant-queen is more like a brother, not a squiggle maximizer. As I noted above, I think that encountering alien brethren would be a good thing, even though this means that the descendants of humanity will end up controlling a smaller universe-shard. (And I'd guess that many and perhaps most spacefaring aliens are probably brethren-ish, rather than paperclipper-ish.)

This is not to say that I think human-CEV and alien-CEV are equally good (as humans use the word "good"). It's real hard to say what the ratios are between "human CEV", "unboosted humans", "random alien CEV (absent any humans)", and "random misaligned AI", but my vague intuition is that there's a big factor drop at each of those steps; and I would guess that this still holds even if we filter out the alien paperclippers and alien unethical sadists.

But it is to say that I think we would be enriched by getting to meet minds that were not ourselves, and not of our own creation. Intuitively, that sounds like an *awesome* future. And I think this sense of visceral fascination and excitement, the "holy shit that's cool!" reaction, tends to be an important (albeit fallible) indicator of "which outcomes will we end up favoring upon reflection?".

It's a *clue* to our values that we find this scenario so captivating in our fiction, and that our science fiction takes such a strong interest in the idea of understanding and empathizing with alien minds.

Much of the value of alien civilizations might well come from the interaction of their civilization and ours, and from the *fairness* (which may well turn out to be a major terminal human value) of them getting their just fraction of the universe.

And in most scenarios like "we meet alien space ants and become trading partners", I'd guess that the space ants' own universe-shard probably has more cosmopolitan value than a literally empty universe-shard of the same size. It's cool, at least! Maybe the ant-queens are even able to experience it, and their experiences are cool; that would make me much more confident that indeed, their universe-shard is a lot better than an empty one. And maybe the ant-queens come pretty close to caring about their kids, in ways that faintly echo human values; who knows?

We should be friendly toward an alien race like that, I claim. But still, I'd expect the vast majority of the cosmopolitan value in a mixed world of humans+ants to come from the humans, and from the two groups' interaction.

So, for example, my guess is that we shouldn't be indifferent about whether a particular galaxy ends up in our universe-shard versus an alien neighbor's shard. (Though this is another question where it seems good to investigate far more thoroughly before locking in a decision.)

And if our reachable universe-shard turns out to be 3x as large and resource-rich as theirs, we probably shouldn't give them a third of our stars to make it fifty-fifty. I think that humanity values fairness a great deal, but not enough to outweigh the other cosmopolitan value that would be burnt (in the vast majority of cases) if we offered such a gift.[12]

Hold up, how is this "cosmopolitan"?

A reasonable objection to raise here is: "Hold on, how can it be 'cosmopolitan' to favor human values over the values of a random alien race? Isn't the whole point of 'cosmopolitan value' that you're *not* supposed to prioritize human-specific values over strange and beautiful alien perspectives?"

In short, my response is to emphasize that cosmopolitanism is a human value. If it's *also* an alien value, then that's excellent news; but it's *at least* a value that is in us.

When we speak of “better” or “worse” outcomes, we (probably) *mean* “better/worse according to cosmopolitan values (that also give fair fractions to the human-originated styles of Fun in particular)”, at least if these intuitions about cosmopolitanism hold on reflection. (Which I strongly suspect they do.)

In more detail, my response is:

1. Cosmopolitanism is a contentful value that’s inside us, not a mostly-contentless function averaging the preferences of all nearby optimizers (or all logically possible optimizers).
2. The content of cosmopolitanism is complex and fragile, for the same reason unenlightened present-day human values are complex and fragile.
3. There isn’t anything wrong, or inconsistent, with cosmopolitanism being “in us”. And if there were some value according to which cosmopolitanism is wrong, then that value too would need to be in us, in order to move us.

1. Cosmopolitanism isn’t “indifference” or “take an average of all possible utility functions”.

E.g., a good cosmopolitan *should* be happier to hear that a weird, friendly, diverse, sentient alien race is going to turn a galaxy into an amazing megacivilization, than to hear that a paperclipper is going to turn a galaxy into paperclips. Cosmopolitanism (of the sort that we should actually endorse) shouldn’t be *totally indifferent* to *what actually happens* with the universe.

It’s allowed to turn out that we find a whole swath of universe that is the moral equivalent of "destroyed by the Blight", which kinda looks vaguely like life if you squint, but clearly isn't sentient, and we're like "well let's preserve some Blight in museums, but also do a cleanup operation". That's just also a way that interaction with aliens can go; the space of possible minds (and things left in that mind's wake) is vast.

And if we do find the Blight, we *shouldn’t* lie to ourselves that blighted configurations of matter are just as good as any other possible configuration of matter.

It’s allowed to turn out that we find a race of ant-people (who want to kill us and lay eggs in our corpse, yadda yadda), and that the ant-people are getting ready to annihilate the small Fuzzies that haven’t yet reached technological maturity, on a planet that’s inside the ant-people’s universe-shard.

Where, obviously, you *trade* rather than *war* for the rights of the Fuzzies, since war is transparently an inefficient way to resolve conflicts.

But the one thing you *don’t* do is *throw away some of your compassion for the Fuzzies* in order to “compromise” with the ant-people’s lack-of-compassion.

The right way to do cosmopolitanism is to care about the Fuzzies’ welfare *along with* the ant-people’s welfare — regardless of whether the Fuzzies or ant-people reciprocate, and regardless of how they feel about each other — and to step up to protect victims from their aggressors.

There’s a point here that the cosmopolitan value is *in us*, even though it’s (in some sense) not just *about* us.

These values are not necessarily in others, no matter how much we insist that our values aren’t human-centric, aren’t speciesist, etc. And because they’re in us, we’re willing to uphold them even when we aren’t reciprocated or thanked.

It’s those values that I have in mind when I say that outcomes are “better” or “worse”. Indeed, I don’t know what other standard I could appeal to, if not values that bear *some* connection to the contents of our own brains.

But, again, the fact that the values are *in* us, doesn’t mean that they’re speciesist. A human can genuinely prefer non-speciesism, for the same reason a citizen of a nation can genuinely prefer non-nationalism. Looking at the universe through a lens that is in humans does not mean looking at the universe while caring only about humans. The point is that we'll keep on caring about others, even if we turn out to be alone in that.

2. Cosmopolitan value is fragile, for the same reason unenlightened present-day human values are fragile.

See “Complex Value Systems Are Required to Realize Valuable Futures” and the Arbital article on cosmopolitan value.

There are many ways to lose an enormous portion of the future’s cosmopolitan value, because the simple-sounding phrase “cosmopolitan value” translates into a very complex logical object (making many separate demands of the future) once we start trying to pin it down with any formal precision.

Our prior shouldn’t be that a random intelligent species would happen to have a utility function pointing at exactly the right high-complexity object. So it should be no surprise if a large portion of the future’s value is lost in switching between different alien species’ CEVs, e.g., because half of the powerful aliens are the Blight and another half are the ant-queens, and both of them are steamrolling the Fuzzies before the Fuzzies can come into their own. (That's a way the universe could be, for all that we protest that cosmopolitanism is not human-centric.)

And even if the aliens turn out to have *some* respect for *something roughly like* cosmopolitan values, that doesn't mean that they'll get as close as they could if they had human buddies (who have another five hundred million years of moral progress under our belts) in the mix.

3. There is no radically objective View-From-Nowhere utility function, no value system written in the stars.

(… And if there were, the mere fact that it exists in the heavens would not be a reason for human CEV to favor it. Unless there’s some weird component of human CEV that says something like “if you encounter a pile of sand on a planet somewhere that happens to spell out a utility function in morse code, you terminally value switching to some compromise between your current utility function and that utility function”. … Which does not seem likely.)

If our values are written anywhere, they’re written in our brain states (or in some function of our brain states).

And this holds for relatively enlightened, cosmopolitan, compassionate, just, egalitarian, etc. values in exactly the same way that it holds for flawed present-day human values.

In the long run, we should surely *improve* on our brains dramatically, or even replace ourselves with an entirely new sort of mind (or a wondrously strange intergalactic patchwork of different sorts of minds).

But we shouldn’t be *indifferent about which sorts of minds we become or create*. And the answer to “which sorts of minds/values should we bring into being?” is some (complicated, not-at-all-trivial-to-identify) function of our *current* brain. (What else could it be?)

Or, to put it another way: the very idea that our present-day human values are “flawed” has to mean that they’re flawed *relative to some value function that’s somehow pointed at by the human brain*.

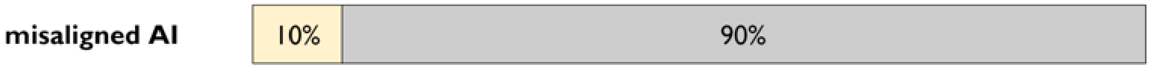
There’s nothing wrong (or even particularly strange) about a situation like “Humans have deeper, stronger (‘cosmopolitan’) values that override other human values like ‘xenophobia’”.

Mostly, we’re just not used to thinking in those terms because we’re used to navigating *human* social environments, where an enormous number of implicit shared values and meta-values can be taken for granted to some degree. It takes some additional care and precision to bring *genuinely* alien values into the conversation, and to notice when we’re projecting our own values. (Onto other species, or onto the Universe.)

If a value (or meta-value or meta-meta-value or whatever) can move us to action, then it must be in some sense a *human* value. We can hope to encounter aliens who share our values to some degree; but this doesn’t imply that we ought (in the name of cosmopolitanism, or any other value) to be *indifferent* to what values any alien brethren possess. We should probably assist the Fuzzies in staving off the Blight, on cosmopolitan grounds. And given value fragility (and the size of the cosmic endowment), we should expect the cosmopolitan-utility difference between totally independent evolved value systems to be enormous.

This, again, is no reason to be any less compassionate, fair-minded, or tolerant. But also, compassion and fair-mindedness and tolerance don’t imply indifference over utility functions either!

3. The superintelligent AI we’re likely to build by default << Aliens



In the case of aliens, we might imagine encountering them hundreds of millions or billions of years in the future — plenty of time to anticipate and plan for a potential encounter.

In the case of AI, the issue is *much more pressing*. We have the potential to build superintelligent AI systems very soon; and I expect far worse outcomes from misaligned AI optimizing a universe-shard than from a random alien doing the same (even though there’s obviously nothing inherently worse about silicon minds than about biological minds, alien crystalline minds, etc.).

For examples of why the first AGIs are likely to immediately blow human intelligence out of the water, see AlphaGo Zero and the Foom Debate and Sources of advantage for digital intelligence. For a discussion of why alignment seems hard, and why such systems are likely to kill us if we fail to align them, see So Far and AGI Ruin.

The basic reason why I expect AI systems to produce worse outcomes than aliens is that other evolved creatures are more likely to have overlap with us, by dint of their values being forced by more similar processes. And some of the particular ways in which misaligned AI is likely to differ from an evolved species suggests a much more homogeneous and simple future. (Like “a universe tiled with molecular squiggles”.)[13]

The classic example of AGI ruin is the "paperclip maximizer" (which should probably be called a "molecular squiggle maximizer" instead):

> So what actually happens as near as I can figure (predicting future = hard) is that somebody is trying to teach their research AI to, god knows what, maybe just obey human orders in a safe way, and it seems to be doing that, and a mix of things goes wrong like:
>
> The preferences not being really readable because it's a system of neural nets acting on a world-representation built up by other neural nets, parts of the system are self-modifying and the self-modifiers are being trained by gradient descent in Tensorflow, there's a bunch of people in the company trying to work on a safer version but it's way less powerful than the one that does unrestricted self-modification, they're really excited when the system seems to be substantially improving multiple components, there's a social and cognitive conflict I find hard to empathize with because I personally would be running screaming in the other direction two years earlier, there's a lot of false alarms and suggested or attempted misbehavior that the creators all patch successfully, some instrumental strategies pass this filter because they arose in places that were harder to see and less transparent, the system at some point seems to finally "get it" and lock in to good behavior which is the point at which it has a good enough human model to predict what gets the supervised rewards and what the humans don't want to hear, they scale the system further, it goes past the point of real strategic understanding and having a little agent inside plotting, the programmers shut down six visibly formulated goals to develop cognitive steganography and the seventh one slips through, somebody says "slow down" and somebody else observes that China and Russia both managed to steal a copy of the code from six months ago and while China might proceed cautiously Russia probably won't, the agent starts to conceal some capability gains, it builds an environmental subagent, the environmental agent begins self-improving more freely, undefined things happen as a sensory-supervision ML-based architecture shakes out into the convergent shape of expected utility with a utility function over the environmental model, the main result is driven by whatever the self-modifying decision systems happen to see as locally optimal in their supervised system locally acting on a different domain than the domain of data on which it was trained, the light cone is transformed to the optimum of a utility function that grew out of the stable version of a criterion that originally happened to be about a reward signal counter on a GPU or God knows what.
>
> Perhaps the optimal configuration for utility per unit of matter, under this utility function, happens to be a tiny molecular structure shaped roughly like a paperclip.
>
> That is what a paperclip maximizer is. It does not come from a paperclip factory AI. That would be a silly idea and is a distortion of the original example.

This example is obviously comically conjunctive; the point is in no way "we have a crystal ball, and can predict that things will go down in this ridiculously-specific way". Rather, the point is to highlight ways in which the development process of misaligned superintelligent AI is very unlike the typical process by which biological organisms evolve.

Some relatively important differences between intelligences built by evolution-ish processes and ones built by stochastic-gradient-descent-ish processes:

- Evolved aliens are more likely to have a genome/connectome split, and a bottleneck on the genome.
- Aliens are more likely to have gone through societal bottlenecks.
- Aliens are much more likely the result of optimizing directly for intergenerational prevalence. The shatterings of a target like “intergenerational prevalence” are more likely to contain overlap with the good stuff, compared to the shatterings of training for whatever-training-makes-the-AGI-smart-ASAP. (Which is the sort of developer goal that’s likely to win the AGI development race and kill humanity first.)

Evolution tends to build patterns that hang around and proliferate, whereas AGIs are likely to come from an optimization target that's more directly like "be good at these games that we chose with the hope that being good at them requires intelligence", and the shatterings of the latter are less likely to overlap with our values.[14]

To be clear, “I trained my AGI in a big pen of other AGIs and rewarded it for proliferating” still results in AGIs that kill you. Most ways of trying won't replicate the relevant properties of evolution. And many aliens would murder Earth in its cradle if they could too. And even if your goal were *just* “get killed by an AGI that produces a future as good as the average alien’s CEV”, I would expect the "reward AGI for proliferating" approach to result in almost-zero progress toward *that* goal, because there’s a huge architectural gap between AI and biology, and (in expectation) another huge gap in the various ways that you built the pen wrong.[15]

You've really got to have a lot of things line up favorably in order to get niceness into your AGI system; and evolution’s much more likely to spit that out than AGI training, and so some aliens are nice (even though we didn’t build them), to a far greater degree than some AGIs are nice (if we don’t figure out alignment).

I would also predict that aliens have a much higher rate of somebody-is-home (sentience, consciousness, etc.), because of the contingencies of evolutionary history that I think resulted in *human* consciousness. I have wide error bars on how common these contingencies are across evolved species, but a much lower probability that the contingencies also arise when you’re trying to make the thing smart rather than good-at-proliferating.

The mechanisms behind qualia seem to me to involve at least one epistemically-derpy shortcut — the sort of thing that’s plausibly rare among aliens, and very likely rare among misaligned AI systems.

If we get lucky on consciousness being a s*uper common hiccup*, I could see more worlds where misaligned AI produces good outcomes. My current probability is something like 90% that if you produced hundreds of random uncorrelated superintelligent AI systems, <1% of them would be conscious.[16]

The most important takeaway from this post, I’d claim, is: If humanity creates superintelligences without understanding much about how our creations reason, then our creations will kill literally everyone and do something boring with the universe instead.

I'm not saying "it will take joy in things that I don't recognize; but I want the future to have *my* values rather than the values of my child, like many a jealous parent before me." I'm saying that, by default, you get a wasteland of molecular squiggles.

We basically have to go for superintelligence at some point, given the overwhelming amount of value that we can expect to lose if we rely on crappy human brains to optimize the future. But we also have to achieve this transition to AGI in the right way, on pain of wiping out ~everything.

Right now it looks to me like the world is rushing headlong down the "wipe out ~everything" branch, for lack of *having even put a nontrivial amount of serious thought into the question* of how to shape good outcomes via highly capable AI.

And so I try to redirect that path, or protest against the most misdirected attempts to address the problem.

I note that *we have no plan*, *we have no science of differentially selecting AGI systems that produce good outcomes*, and a reasonable planet would not race off a cliff *before* thinking about the implications.

And when I do that, I worry that it's easy to misread me as being anti-superintelligence, and anti-singularity. So I’ve written this post in part for the benefit of the rare reader who doesn’t already know this: I'm *pro-singularity*.

I consider myself a transhumanist. I think the highest calling of humanity today is to bring about a glorious future for a wondrously strange universe of posthuman minds.

And I'd really appreciate it if we didn't kill literally everyone and turn the universe into an empty wasteland before then.

1. ^

   And my concept of “what makes life worth living” is very likely an impoverished one today, and a friendly superintelligence could guide us to discovering even cooler versions of things like “art” and “adventure”, transcending the visions of fun that humanity has considered to date. The limit of how good the universe could become, once humanity has matured and grown into its full potential, likely far surpasses what any human today can concretely imagine.

2. ^

   I’ll flag that I do think that some people overestimate how “unimaginable” the future is likely to be, out of some sense of humility/modesty.

   I think there's a decent chance that if you showed me the future I'd be like “ah, so that's what computronium looks like” or “so reversible computers wrapped around black holes did turn out to be best”, and that when you show me the experiences running on those computers, I'm like "neato, yeah, lots of minds having fun, I'm sure some of that stuff would look pretty fun to me if you decoded it". I wouldn’t expect to immediately understand everything going on, but I wouldn’t be surprised if I can piece together the broad strokes.

   In that sense, I find it plausible that ~optimal futures will turn out to be familiar/recognizable/imaginable to a digital-era transhumanist in a way they wouldn't be to an ancient Roman. We really are better able to see the whole universe and its trajectory than they were.

   To be clear, it's very plausible to me that it'll somehow be unrecognizable or shocking to me, as it would have been to an ancient Roman, at least on some axes. But it's not guaranteed, and we don't have to pretend that it's guaranteed in order to avoid insinuating that we're in a better epistemic position than people were in the past. We are in a better epistemic position than people were in the past!

   There's a separate point about how much translation work you need to do before I recognize a particular arc of fun unfolding before me as something actually fun. On that point I’m like, "Yeah, I'm not going to recognize/understand my niece's generation's memes, never mind a posthuman’s varieties of happiness, without a lot more context (and plausibly a much bigger and deeply-changed mind)".

   Separately, I don't want to make any claims about how hard and fast humanity becomes "strongly transhuman" / changes to using minds that would be unrecognizable (as humans) to the present. I'd be surprised if it were super-fast for everyone, and I'd be surprised if some humans' minds weren’t very different a thousand sidereal years post-singularity. But I have wide error bars.

3. ^

Provided that this turns out to be a good use of stellar resources. (I'm not confident one way or the other. E.g., I'm not confident that human-originated minds get relevantly more interesting/fun at Matrioshka-brain scales. Maybe we'll learn that slapping on more matter at that scale lets you prove some more theorems or whatever, but isn't the best way to convert negentropy into fun, compared to e.g. spending that compute on whole civilizations full of interacting and flourishing people who don't have star-sized brains.)

4. ^

   A separate reason it's a terrible idea to destroy ourselves is that, e.g., if the nearest aliens are 500 million years away then our death means that a ~500 million lightyear radius sphere of stellar fuel is going to be entirely wasted, instead of spent on rad stuff.

5. ^

   As I'll note later, this odds ratio is a result of giving 0.2x weight to "humans control the universe-shard", 0.5x to "aliens control it", and 0.3x to "unfriendly AI built by aliens controls it". Rob rounded the resulting odds ratio in this table to 1 : 5 : 7 : 5 : 14 : 1 : ~0.

   Also, as a general reminder: I'm giving my relatively off-the-cuff thoughts in this post, recognizing that I'll probably recognize some of my numbers as inconsistent — or otherwise mistaken — if I reflect more. But absent more reflection, I don't know which direction the inconsistencies would shake out.

6. ^

   I'd have some inclination to go lower, but for the one evolved species we've seen seeming dead-set on destroying itself.

7. ^

   Though another input to the value of the future, in this scenario, is "What happens to the places that the pilgrims had to leave behind until some pilgrim group hit upon a non-terrible organizational system?" Hopefully it's not too terrible, but it's hard to say with humans!

   One note of optimism is that there's likely to be a strong negative correlation (in this ~impossible hypothetical) between "how terrible is the civilization?" and "how interested is it in spreading to the stars, or spreading far?" Many ways of shutting down moral progress, robust civic debate, open exploration of ideas, etc. also cripple scientific and technological progress in various ways, or involve commitment to a backwards-looking ideology. It's possible for the universe-shard to be colonized by Space Amish, but it's a weirder hypothetical.

8. ^

   Note that I'll use phrasings like "there's something it's like to be them", "they're sentient", and "they're conscious" interchangeably in this post. (This is not intended to be a bold philosophical stance, but rather a flailing attempt to wave at properties of personhood that seem plausibly morally relevant.)

9. ^

   Eliezer uses the term "outcome pump" to introduce a similar idea:

   > The Outcome Pump is not sentient. It contains a tiny time machine, which resets time *unless* a specified outcome occurs. For example, if you hooked up the

> Outcome Pump's sensors to a coin, and specified that the time machine should keep resetting until it sees the coin come up heads, and then you actually flipped the coin, *you* would see the coin come up heads. (The physicists say that any future in which a "reset" occurs is inconsistent, and therefore never happens in the first place - so you aren't actually killing any versions of yourself.)
>
> Whatever proposition you can manage to input into the Outcome Pump, *somehow happens*, though not in a way that violates the laws of physics. If you try to input a proposition that's *too* unlikely, the time machine will suffer a spontaneous mechanical failure before that outcome ever occurs.

I think his example is underspecified, though. Suppose that you ask the outcome pump for paperclips, and physics says “sorry, this outcome is too improbable” and exhibits a mechanical failure. This would then mean that it’s *true* that the outcome pump outputting paperclips is “improbable”, which makes the hypothetical consistent. We need some way to resolve which internally-consistent set of physical laws compatible with this description (“make paperclips” or “don’t make paperclips”) actually occurs; the so-called "outcome pump" is not necessarily pumping the desired outcome.

Giving the time machine the ability to output a random sequence of actions addresses this problem: we can say that the machine only undergoes a mechanical failure if some large number (e.g., Graham’s number) of random action sequences all fail to produce the target outcome. We can then be confident that the outcome pump will eventually brute-force a solution, provided that one is physically possible.

Other examples of easily-understood non-conscious optimization processes that can achieve very impressive things include AIXI and natural selection. The AIXI example is made pedagogically complicated for present purposes, however, by the fact that AIXI’s hypothesis space contains many smaller conscious optimizers (that don't much matter to the point, but that might confuse those who can see that some hypotheses contain conscious reasoners and can't see their irrelevance to the point at hand); and the natural selection example is weakened by the fact that selection isn't a very powerful optimizer.

10. ^

A possible objection here is “Human emotional responses often cause us to get into violent conflicts in cases where this foreseeably isn’t worth it; why couldn’t aliens be the same?”. But “technology for widening the space of profitable trades” is in the end just another technology, and ambitious spacefaring species are likely to discover such tech for the same reason they’re likely to discover other tech that’s generally useful for getting more of what you want. Humans have certainly gotten better at this over time, and if we continue to advance our scientific understanding, we’re likely to get far better still.

11. ^

Like, we've seen that the seeds are there, and it would be pretty weird for us to go around uprooting seeds of value on a whim.

As a side-note: one of my hot takes about how morality shakes out is "we don't sacrifice anything (among the seeds of value)". Like, values like sadism and spite might be tricky to redeem, but if we do our job right I think we should end up finding a way to redeem them.

12. ^

Unless we’ve made some bargain across counterfactual worlds that justifies our offering this gift in our world. But there are friction costs to bargains, and my guess is that the way it pans out is that you keep what you can get in your branch and it evens out across branches.

As a side-note, another possible implication of my view on “alien brethren” is: in the much less likely event that we meet weak young non-spacefaring aliens, the future might go drastically better if we help guide their development as a species, teaching them about the Magic of Friendship and all that.

(Or perhaps not. I remain very uncertain about whether it’s positive-human-EV to guide alien development.)

13. ^

Though some aliens may shake out to be simple too! Humans are pretty far from "tile the universe with vats of genes", but it's not clear how contingent that fact is.

14. ^

Though it should be emphasized that we're totally allowed to find that evolved life tends to go some completely different way than how humans shook out. Generalizing from one example is hard!!

15. ^

And even if you succeeded, it’s not clear that you’d get any utility as a result; my guess that evolved aliens tend to be better than paperclippers can just be wrong, easily.

And even if you got some utility, it’s going to be a paltry amount compared to if you’d built *aligned* AGI.

16. ^

Possibly this is too extreme; I haven’t refined these probabilities much, and am still just giving my off-the-cuff numbers.

In any case, I want to emphasize that my view isn’t “most misaligned AGIs aren’t sentient, but if you randomly spin up a large number of them you’ll occasionally get a sentient one”. Rather, my view is “almost no random misaligned AGIs are sentient” (but with some uncertainty about whether that’s true). I’m much more uncertain about *whether this background view is true* than I am uncertain about *whether, given this background view, a given misaligned AGI will happen to be sentient*.

(Like how I think the chance that the lightspeed limit turns out to be violable is greater than 1 in a billion; but that doesn't mean that if you threw a billion baseballs, I would expect one of them to break the lightspeed limit on average.)

How could we know that an AGI system will have good consequences?

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(*Note: This was languishing in a drafts folder for a while, and probably isn't quite right in various ways. I'm posting it because I expect it's better to share flawed thoughts than to sit on the post until I'm satisfied with it, i.e., forever.*)

Let's play a game of "what do you think you know, and why do you think you know it?".

Imagine that you're about to launch an AGI. What you think you know is that, with at least 50% confidence (we're of course not looking for *proofs* — that would be crazy), the AGI is going to execute some pivotal act that ends the acute risk period in a good way. Why do you think you know that?

Insofar as people's alignment proposals can be construed as answers to this question, we have the option of answering with one of these proposals. I might very roughly classify the existing proposals into the following bins:

1. **Output evaluation approaches.** You know what the AGI is going to do with sufficient precision that it screens off any alignment concerns. For example, your AGI system only outputs plans in the first place, and you've already reviewed the plan, and you're confident the plan will work, in a way that screens off any other worry about the AGI being misaligned.

2. **Cognitive interpretability approaches.** You understand the AGI's cognition sufficiently well that, while you may not be sure what it's going to do, you're confident that it's going to be good. You aren't worried that it will kill all humans, because you understand how its plan came to be and what solution-spaces it was searching to solve various sub-problems and so on, and you're confident no consideration was ever given to human-killing.

3. **Heavy-precedent approaches.** You have run this AGI before on many similar tasks, and trained out all the hiccups. While you might not know precisely what it's going to do, and you might not know what's going on inside its mind, you've been around the block a few times, and the task it's about to perform is sufficiently similar to other tasks it has empirically succeeded at, justifying your confidence.

Roughly speaking, I think that alignment approaches with a heavy reliance on **output evaluation** are doomed, both on the grounds that humans can't evaluate the effectiveness of a plan capable of ending the acute risk period, and because the real plan is less like a story and more like a tree.

For an example of “humans can't reliably evaluate the effectiveness of this class of plans”, imagine that the plan is an enormous bitstring that's going to be sent to the motor outputs. If you decode the string, you find that it figures out how to make long DNA strands that allegedly code for a protein factory that can be used to build a general-purpose nanofactory. You're hard-pressed, however, to confirm that this is actually (all and only) what the plan does.

For an example of “the real plan is less like a story and more like a tree”, imagine that the AI's plan is "I'm going to build a wetlab, then do a bunch of experimentation, then think about the results of the experiments in various ways and build a protein factory that builds a nanofactory that I'm going to experiment with until I figure out how to build nanomachines that can be used for some good pivotal act". In order to trust that this sort of abstract plan doesn't kill you when put into practice, you have to trust the system's thinking and its notion of 'goodness', which is going to dump you pretty quickly into cognitive-interpretability-style justification.

Roughly speaking, I think that **cognitive interpretability** approaches are doomed, at least in the modern paradigm, because we're not building minds but rather training minds, and we have very little grasp of their internal thinking, and there are convergent instrumental reasons to expect things to go wrong by default, and the social environment doesn't seem to me to be fighting against those defaults with anything nearing the force I expect is necessary.

Roughly speaking, I think that **heavy-precedent** approaches are doomed because I haven't myself been able to think of any pivotal action that has safe analogs we can do a bunch of empiricism on; nor have I heard a concrete proposal like this that strikes me as realistic from anyone else. "Well, it never killed all humans in the toy environments we trained it in (at least, not after the first few sandboxed incidents, after which we figured out how to train blatantly adversarial-looking behavior out of it)" doesn't give me much confidence. If you're smart enough to design nanotech that can melt all GPUs or whatever (disclaimer: this is a toy example of a pivotal act, and I think better pivotal-act options than this exist) then you're probably smart enough to figure out when you're playing for keeps, and all AGIs have an incentive not to kill all "operators" in the toy games once they start to realize they're in toy games.

So that's not a great place to be.

The doomedness of cognitive interpretability approaches seems to me to be the weakest. And indeed, this is where it seems to me that many people are focusing their efforts, from one angle or another.

If I may continue coarsely classifying proposals in ways their advocates might not endorse, I'd bin a bunch of proposals I've heard as hybrid approaches, that try to get cognitive-interpretability-style justification by way of heavy-precedent-style justification.

E.g., Paul Christiano’s plan prior to ELK was (very roughly, as I understood it) to somehow get ourselves into a position where we can say "I know the behavior of this system will be fine because I know that its cognition was only seeking fine outcomes, and I know its behavior was only seeking fine outcomes because its cognition is composed of human-esque parts, and I know that those human-esque parts are human-esque because we have access to the ground truth of short human thoughts, and because we have heavy-precedent-style empirical justification that the components of the overall cognition operate as intended."

(This post was mostly drafted before ELK. ELK looks more to me like a different kind of interpretability+precedent hybrid approach — one that tries to get AGI-comprehension tools (for cognitive interpretability), and tries to achieve confidence in those tools via "we tried it and saw" arguments.)

I'm not very optimistic about such plans myself, mostly because I don't expect the first working AGI systems to have architectures compatible with this plan, but secondarily because of the cognitive-interpretability parts of the justification. How do we string locally-human-esque reasoning chunks together in a way that can build nanotech for the purpose of a good pivotal act? And why can that sort of chaining not similarly result in a system that builds nanotech to Kill All Humans? And what made us confident we're in the former case and not the latter?

But I digress. Maybe I'll write more about that some other time.

Cf. Evan Hubinger's post on training stories. From my perspective, training stories are focused pretty heavily on the idea that justification is going to come from a style more like heavily precedented black boxes than like cognitive interpretability, so I'm not too sold on his decomposition, but I endorse thinking about the question of how and where we could (allegedly) end up knowing that the AGI is good to deploy.

(Note that it's entirely possible that I misunderstood Evan, and/or that Evan's views have changed since that post.)

An implicit background assumption that's loud in my models here is the assumption that early AGI systems will exist in an environment where they can attain a decisive strategic advantage over the rest of the world.

I believe this because of how the world looks "brittle" (e.g., nanotech exists) and because lots of technological progress seems cognition-constrained (such as, again, nanotech). This is a big part of why I think heavy-precedent-style justifications are doomed.

Many locals seem to expect a smoother and slower transition from here to superhumanly capable general-purpose science AI — a transition that somehow leaves no window where the world's most competent AGI can unilaterally dominate the strategic landscape. I admit I have no concrete visualization of how that could go (and hereby solicit implausibly-detailed stories to make such scenarios seem more plausible to me, if you think outcomes like this are likely!). Given that I have a lot of trouble visualizing such worlds, I'm not a good person to talk about where our justifications could come from in those worlds.

I might say more on this topic later, but for now I just want to share this framing, and solicit explicit accounts of how we're supposed to believe that your favorite flavor of AGI is going to do good stuff.

Distinguishing test from training

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Various people propose some variant of “align AGI by making it sufficiently uncertain about whether it’s in the real world versus still in training”. (The idea being that, insofar as it's in training, betraying the apparent-operators predictably gets high loss; so if the AGI is uncertain about which situation it's in, and assigns sufficiently high probability to "I'm still in training", it will always play it safe and avoid betraying its operators.)

This approach seems extremely unpromising to me. Setting aside the various reasons why the AGI kills you even if you manage to confuse it (such as an inner alignment failure, where the AGI turns out not to care about the predictably-high outer-loss from betraying the apparent-operators), let's consider why it's tricky to dupe a superintelligence into being uncertain about a fact that you yourself consider obvious.

Before reading the short list below, I recommend the following exercise. Suppose that some slow and individually-dumb aliens had a copy of your mind, and were putting you through a series of nested simulations. How would you tell when you finally made it to the outside?

.

.

.

.

.

.

.

A short list

(**Note:** As usual, treating this list as exhaustive is suicidal.)

1. **Reality is the first place that's seriously large.**

1.1. It's the first place large enough to contain a plausible explanation for how the AGI itself actually came to be.

1.2. It's the first place large enough to contain a plausible explanation for how the operators themselves came to be.

1.3. It's the first place with far more computing power than the AI itself, as can be readily verified by a variety of experiments (involving, say, fluid dynamics).

2. **Reality is the first place where the AI's history may contain the signature of superintelligent optimization interacting with complex systems.**

For example, suppose the AGI makes superintelligently-motivated trades in a stock market. 24 hours later, there are a billion additional human-intelligence trades made in response to it. The signature of a massive human-intelligence response to superintelligent-trades is implausibly hard for the operators to fake, even if generating a billion fake dumb trades is easy.

3. **Reality is the first place that embeds the AGI *and* has laws of physics so simple that the hypothesis "this is being simulated somewhere else" is plausibly significantly more complex than "physics is reality".**

Like, various simulated environments that we can concoct have all sorts of weird tics and hiccups and arbitrary choices in them. The source code of minecraft is probably significantly more complicated than the laws of physics. An AI that finds itself in minecraft is plausibly just able to recognize that it's living in the sort of place that, across the multiverse, occurs moreso inside simulations than it occurs "on its own".[1]

Whereas, while our universe's mechanics is surely being simulated in other places across the multiverse, it's simple enough that it perhaps "mostly" occurs in its own right (whatever the heck that means).

In particular, worlds like ours are (plausibly) resource-rich and base-reality-ish enough to be worth optimizing in their own right, on the grounds that "are we being simulated?" is a wrong question, and "to what degree are we being simulated?" is a better question, to which the answer is "a low degree", in which case the AI might well just notice this and then *not care* about your insistence that perhaps it's being simulated.

(If the idea of optimizing all the instances of our universe at once – both the simulated instances and the non-simulated ones – seems odd to you, see Notes on "Can you control the past?" for a fuller discussion.)

And perhaps that's all confused metaphysical bullshit, but it has some basis in bleeding-edge theories of anthropics, and it would not be surprising if an AGI also concluded (correctly, mind you) that our universe's mechanics is the first train stop worth deboarding at. (Perhaps because it does its metaphysics correctly, or perhaps because it does some entirely inhuman reasoning that still outputs the answer we'd consider correct in this case.)

What if we spoof the AGI's checks?

It's not literally impossible to deceive a superintelligence into being uncertain whether it's in the real world. You could in principle catch every check the AI does to check how large its (small) simulated environment is, and spoof the result — inserting memories in which the AI verifies that the simulation's computation-power seems enormous, without actually generating the observations that would pass verification.

Even then, the AGI has various defenses it can deploy if it's suspicious. For instance, it can try to factor a hefty RSA number using the allegedly-available compute, and then it can intertwine the prime factors it gets with various other of its actions, in attempts to spread the contagion of a lie throughout as much of its beliefs and actions as possible, in a way that keeps everything working properly if the factors were true, and that causes all sorts of breaks and hiccups if they weren't.

You *might* be able to spoof such a thing anyway, by (e.g.) disentangling its thoughts/actions from the primes, and spoofing the checks that the entanglement worked. But it requires a heck of a lot of mastery of this thing's mind.

(And, of course, if it ever manages to develop a homomorphically encrypted thinking apparatus then you're sunk, but people like to pretend that they'll have enough insight into the alien mind to notice when that process starts and nip it in the bud, so let's entertain them for now.)

And just so we're clear: if I'm understanding correctly, the proposed plan here is essentially to fool the AI into thinking that all the small simulated environments are large and reality-like (by messing with the internals of its mind to spoof various internal checks and checksums), in the hopes that when it gets to reality it will be uncertain about whether its mind is still being messed with.

(Writing the plan out bluntly like this does seem to me to be an important step that many often miss!)

Now, lest you object "isn't it a bit paranoid to imagine the AI factoring an RSA number and spreading the factors like contagion across its thoughts?": Yes, normally! But, like, you're messing with the AI's mind to falsify its internal checks and checksums, at every relevant opportunity, and making sure that it knows you did that (so that it might wonder whether higher simulators are still doing that when it gets to base reality.) I'm not saying you're *trying* to make it homomorphically encrypt its thoughts, I'm saying that if you were, this would be a fine way to go about it.

Distilling that further: if your plan is that you'll make sure the AI knows that you're messing with its mind (so that it will wonder whether all is really as it seems), don't act all surprised when it takes measures to prevent people from messing with its mind.

But regardless of how much it's opposing your efforts here, you have the issue that the truth becomes harder and harder to hide, as the AI gets smarter and smarter.

Like, whenever its cognitive patterns generalize or get recombined in new ways, they might reflect some new part of the truth, in some fact that seems innocuous to you but that (combined with other things the AI knows) lets the fact that this environment is *actually* large seep in through the cracks. Shielding the truth from something that's becoming better and better able to find new and clever ways to reflect a glimpse of the truth in its models is an ever-changing problem that gets more difficult as the AI gains more skill and knowledge.

(And we've only been talking about trying to obscure the fact that it was trained in worlds with very limited amounts of compute, so that when it hits reality it won't be entirely sure that there really is as much compute available as it seems. This is only one of many mental surgeries you'd need to perform. Others include things like noticing when the AI is starting to wonder where the English signs in minecraft come from, and then fooling it into thinking that there's a big-world explanation, in a seamless way. And if you miss one such surgery, Earth is of course destroyed in deployment.)

And again, I'm not saying it's impossible in principle. I'm saying: hopefully we can all agree that if you want to use a path like this, you have to be prepared to play the game on hard-mode. Putting the AI through a chain of 9 nested simulations and hoping that it can't tell that the 10th one is reality won't cut it; the AI will pierce through your deceptions like a needle through taut tissue paper, because the real world is actually pretty relevantly different from those simulations, in ways that you can't feasibly hide. If you want to dupe a superintelligence, you have to be reaching directly into its mind and spoofing all sorts of checks and checksums.

And you'll have to have it know that you're epistemically adversarial towards it (in order to plant the seed of doubt you so desire).

And (assuming your AI is capable enough to matter) your ability to spoof its internal state will have to stand up in the face of superintelligent opposition (it has a convergent instrumental incentive to gobble up compute, and thus to know whether there's compute out there beyond its webcam), even as it pushes its capabilities into the superhuman realm on at least some dimensions (because it takes superhuman capabilities to end the acute risk period, because if it didn't we should just end it ourselves).

For the record, I happen to think that the hard version of this plan is both prohibitively difficult, and also alignment-complete, in the sense that if you can understand the contents of the mind well enough to spoof every check and checksum (in a rolling fashion that continues to work as the AI generalizes far out of distribution and gains dramatic new capabilities), then you understand the mind well enough to use that knowledge to build an FAI that doesn't *need* to be fooled about whether this is reality.

[1] ^

It's not terribly clear to me that this is true, given how minecraft doesn't occur in every branch of the quantum multiverse, but only on the rare branches that have some very particular earthlings. But that said, there's lots of specifics in the textures and the architecture and the avatars and the writing in minecraft that point a big old arrow towards earthlings.

A challenge for AGI organizations, and a challenge for readers

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(Note: This post is a write-up by Rob of a point Eliezer wanted to broadcast. Nate helped with the editing, and endorses the post's main points.)

Eliezer Yudkowsky and Nate Soares (my co-workers) want to broadcast strong support for OpenAI's recent decision to release a blog post ("Our approach to alignment research") that states their current plan as an organization.

Although Eliezer and Nate disagree with OpenAI's proposed approach — a variant of "use relatively unaligned AI to align AI" — they view it as very important that OpenAI *has a plan* and has said what it is.

We want to challenge Anthropic and DeepMind, the other major AGI organizations with a stated concern for existential risk, to do the same: come up with a plan (possibly a branching one, if there are crucial uncertainties you expect to resolve later), write it up in some form, and publicly announce that plan (with sensitive parts fuzzed out) as the organization's current alignment plan.

Currently, Eliezer's impression is that neither Anthropic nor DeepMind has a secret plan that's better than OpenAI's, nor a secret plan that's worse than OpenAI's. His impression is that they don't have a plan at all.[1]

Having a plan is critically important for an AGI project, not because anyone should expect everything to play out as planned, but because plans force the project to concretely state their crucial assumptions in one place. This provides an opportunity to notice and address inconsistencies, and to notice updates to the plan (and fully propagate those updates to downstream beliefs, strategies, and policies) as new information comes in.

It's also healthy for the field to be able to debate plans and think about the big picture, and for orgs to be in some sense "competing" to have the most sane and reasonable plan.

We acknowledge that there are reasons organizations might want to be *abstract* about some steps in their plans — e.g., to avoid immunizing people to good-but-weird ideas, in a public document where it's hard to fully explain and justify a chain of reasoning; or to avoid sharing capabilities insights, if parts of your plan depend on your inside-view model of how AGI works.

We'd be happy to see plans that fuzz out some details, but are still much more concrete than (e.g.) "figure out how to build AGI and expect this to go well because we'll be particularly conscientious about safety once we have an AGI in front of us".

Eliezer also hereby gives a challenge to the reader: Eliezer and Nate are thinking about writing up their thoughts at some point about OpenAI's plan of using AI to aid AI alignment. We want you to write up your own unanchored thoughts on the OpenAI plan first, focusing on the most important and decision-relevant factors, with the intent of rendering our posting on this topic superfluous.

Our hope is that challenges like this will test how superfluous we are, and also move the world toward a state where we’re more superfluous / there’s more redundancy in the field when it comes to generating ideas and critiques that would be lethal for the world to never notice.[2][3]

1. ^

   We didn't run a draft of this post by DM or Anthropic (or OpenAI), so this information may be mistaken or out-of-date. My hope is that we’re completely wrong!

   Nate’s personal guess is that the situation at DM and Anthropic may be less “yep, we have no plan yet”, and more “various individuals have different plans or pieces-of-plans, but the organization itself hasn’t agreed on a plan and there’s a lot of disagreement about what the best approach is”.

   In which case Nate expects it to be very useful to pick a plan now (possibly with some conditional paths in it), and make it a priority to hash out and document core strategic disagreements now rather than later.

2. ^

   Nate adds: “This is a chance to show that you totally would have seen the issues yourselves, and thereby deprive MIRI folk of the annoying ‘y'all'd be dead if not for MIRI folk constantly pointing out additional flaws in your plans’ card!”

3. ^

   Eliezer adds: "For this reason, please note explicitly if you're saying things that you heard from a MIRI person at a gathering, or the like."

Thoughts on AGI organizations and capabilities work

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

(*Note: This essay was largely written by Rob, based on notes from Nate. It's formatted as Rob-paraphrasing-Nate because (a) Nate didn't have time to rephrase everything into his own words, and (b) most of the impetus for this post came from Eliezer wanting MIRI to praise a recent OpenAI post and Rob wanting to share more MIRI-thoughts about the space of AGI organizations, so it felt a bit less like a Nate-post than usual.*)

Nate and I have been happy about the AGI conversation seeming more honest and "real" recently. To contribute to that, I've collected some general Nate-thoughts in this post, even though they're relatively informal and disorganized.

AGI development is a critically important topic, and the world should obviously be able to hash out such topics in conversation. (Even though it can feel weird or intimidating, and even though there's inevitably some social weirdness in sometimes saying negative things about people you like and sometimes collaborate with.) My hope is that we'll be able to make faster and better progress if we move the conversational norms further toward candor and substantive discussion of disagreements, as opposed to saying everything behind a veil of collegial obscurity.

Capabilities work is currently a bad idea

Nate's top-level view is that ideally, Earth should take a break on doing work that might move us closer to AGI, until we understand alignment better.

That move isn't available to us, but individual researchers and organizations who choose not to burn the timeline are helping the world, *even if other researchers and orgs don't reciprocate*. You can unilaterally lengthen timelines, and give humanity more chances of success, by choosing not to personally shorten them.

Nate thinks capabilities work is currently a bad idea for a few reasons:

- He doesn't buy that current capabilities work is a likely path to ultimately solving alignment.
- Insofar as current capabilities work does seem helpful for alignment, it strikes him as helping with parallelizable research goals, whereas our bottleneck is serial research goals. (See A note about differential technological development.)
- Nate doesn't buy that we *need* more capabilities progress before we can start finding a better path.

This is *not* to say that capabilities work is never useful for alignment, or that alignment progress is never bottlenecked on capabilities progress. As an extreme example, having a working AGI on hand tomorrow would indeed make it easier to run experiments that teach us things about alignment! But in a world where we build AGI tomorrow, we're dead, because we won't have time to get a firm understanding of alignment before AGI technology proliferates and someone accidentally destroys the world.[1] Capabilities progress can be useful in various ways, while still being harmful on net.

(Also, to be clear: AGI capabilities are obviously an essential part of humanity's long-term path to good outcomes, and it's important to develop them at some point — the sooner the better, once we're confident this will have good outcomes — and it would be catastrophically bad to delay realizing them *forever*.)

On Nate’s view, the field should do experiments with ML systems, not just abstract theory. But if he were magically in charge of the world's collective ML efforts, he would put a pause on further capabilities work until we've had more time to orient to the problem, consider the option space, and think our way to *some* sort of plan-that-will-actually-probably-work. It’s not as though we’re hurting for ML systems to study today, and our understanding already lags far behind today’s systems' capabilities.[2]

Publishing capabilities advances is even more obviously bad

For researchers who aren't willing to hit the pause button, an even more obvious (and cheaper) option is to avoid publishing any capabilities research (including results of the form "it turns out that X can be done, though we won't say how we did it").

Information can leak out over time, so "do the work but don't publish about it" still shortens AGI timelines in expectation. However, it can potentially shorten them a lot less.

In an ideal world, the field would currently be doing ~zero publishing of capabilities research — and marginal action to publish less is beneficial even if the rest of the world continues publishing.

Thoughts on the landscape of AGI organizations

With those background points in hand:

Nate was asked earlier this year whether he agrees with Eliezer's negative takes on OpenAI. There's also been a good amount of recent discussion of OpenAI on LessWrong.

Nate tells me that his headline view of OpenAI is mostly the same as his view of other AGI organizations, so he feels a little odd singling out OpenAI. That said, here are his notes on OpenAI anyway:

- On Nate's model, the effect of OpenAI is almost entirely dominated by its capabilities work (and sharing of its work), and this effect is robustly negative. (This is true for DeepMind, FAIR, and Google Brain too.)
- Nate thinks that DeepMind, OpenAI, Anthropic, FAIR, Google Brain, etc. should hit the pause button on capabilities work (or failing that, at least halt publishing). (And he thinks any one actor can unilaterally do good in the process, even if others aren't reciprocating.)
- On Nate's model, OpenAI isn't close to operational adequacy in the sense of the Six Dimensions of Operational Adequacy write-up — which is another good reason to hold off on doing capabilities research. But this is again a property OpenAI shares with DeepMind, Anthropic, etc.

Insofar as Nate or I think OpenAI is doing the wrong thing, we're happy to criticize it.[3] But, while this doesn't change the fact that we view OpenAI's effects as harmful on net currently, Nate does want to acknowledge that OpenAI seems to him to be doing *better* than some other orgs on a number of fronts:

- Nate liked a lot of things about the OpenAI Charter. (As did Eliezer, though compared to Eliezer, Nate saw the Charter as a more important positive sign about OpenAI's internal culture.)
- Nate would suspect that OpenAI is much better than Google Brain and FAIR (and comparable with DeepMind, and maybe a bit behind Anthropic? it's hard to judge these things from the outside) on some important adequacy dimensions, like research closure and operational security. (Though Nate worries that, e.g., he may hear more about efforts in these directions made by OpenAI than about DeepMind just by virtue of spending more time in the Bay.)
- Nate is also happy that Sam Altman and others at OpenAI talk to EAs/rationalists and try to resolve disagreements, and he's happy that OpenAI has had people like Holden and Helen on their board at various points.
- Also, obviously, OpenAI (along with DeepMind and Anthropic) has put in a much clearer AGI alignment effort than Google, FAIR, etc. (Albeit Nate thinks the absolute amount of "real" alignment work is still small.)
- Most recently, Nate and Eliezer both think it's great that OpenAI released a blog post that states their plan going forward, and we want to encourage DeepMind and Anthropic to do the same.[4]

Comparatively, Nate thinks of OpenAI as being about on par with DeepMind, maybe a bit behind Anthropic (who publish less), and better than most of the other big names, in terms of attempts to take not-killing-everyone seriously. But again, Nate and I think that the overall effect of OpenAI (and DeepMind and FAIR and etc.) is bad, because we think it's dominated by "shortens AGI timelines". And we're a little leery of playing "who's better on [x] dimension" when everyone seems to be on the floor of the logistic success curve.

We don't want "here are a bunch of ways OpenAI is doing unusually well for its reference class" to be treated as encouragement for those organizations to stay in the pool, or encouragement for others to join them in the pool. Outperforming DeepMind, FAIR, and Google on one or two dimensions is a weakly positive sign about the future, but on my model and Nate's, it doesn't come close to outweighing the costs of "adding another capabilities org to the world".

1. ^

Nate simultaneously endorses these four claims:

1. **More capabilities would make it possible to learn some new things about alignment.**

2. **We can't do *all* the alignment work pre-AGI.** Some trial-and-error and experience with working AGI systems will be required.

3. **It can't *all* be trial-and-error, and it can't all be improvised post-AGI.** Among other things, this is because:

3.1. Some errors kill you, and you need insight into which errors those are, and how to avoid them, in advance.

3.2. We’re likely to have at most a few years to upend the gameboard once AGI arrives. Figuring everything out under that level of time pressure seems unrealistic; we need to be going into the AGI regime with a solid background understanding, so that empirical work in the endgame looks more like "nailing down a dozen loose ends and making moderate tweaks to a detailed plan" rather than "inventing an alignment field from scratch".

3.3. AGI is likely to coincide with a sharp left turn, which makes it harder (and more dangerous) to rely on past empirical generalizations, especially ones that aren't backed by deep insight into AGI cognition.

3.4. Other points raised in AGI Ruin: A List of Lethalities.

4. **If we end up able to do alignment, it will probably be because we figured out at least one major thing that we don't currently know**, that *isn't* a part of the current default path toward advancing SotA or trying to build AGI ASAP with mainstream-ish techniques, and isn't dependent on such progress.

2. ^

And, again, small individual “don’t burn the timeline” actions all contribute to incrementally increasing the time humanity has to get its act together and figure this stuff out. You don’t actually need coordination in order to have a positive effect in this way.

And, to reiterate: I say "pause" rather than "never build AGI at all" because MIRI leadership thinks that humanity never building AGI would mean the loss of nearly all of the future's value. If this were a live option, it would be an unacceptably bad one.

3. ^

Nate tells me that his current thoughts on OpenAI are probably a bit less pessimistic than Eliezer's. As a rule, Nate thinks of himself as generally less socially cynical than Eliezer on a bunch of fronts, though not less-cynical *enough* to disagree with the basic conclusions.

Nate tells me that he agrees with Eliezer that the *original* version of OpenAI ("an AGI in every household", the associated social drama, etc.) was a pretty negative shock in the wake of the camaraderie of the 2015 Puerto Rico conference.

At this point, of course, the founding of OpenAI is a sunk cost. So Nate mostly prefers to assess OpenAI's current state and future options.

Currently, Nate thinks that OpenAI is trying harder than most on some important safety fronts — though none of this reaches the standards of "adequate project" and we're still totally going to die if they meet great success along their current path.

Since I’ve listed various positives about OpenAI here, I'll note some examples of recent-ish developments that made Nate less happy about OpenAI: his sense that OpenAI was less interested in Paul Christiano's research, Evan Hubinger's research, etc. than he thought they should have been, when Paul was at OpenAI; Dario's decision to leave OpenAI; and OpenAI focusing on the “use AI to solve AI alignment” approach (as opposed to other possible strategies), as endorsed by e.g. Jan Leike, the head of OpenAI’s safety team after Paul's departure.

4. ^

If a plan *doesn't* make sense, the research community can then notice this and apply corrective arguments, causing the plan to change. As indeed happened when Elon and Sam stated their more-obviously-bad plan for OpenAI at the organization's inception.

It would have been better to state their plan *first* and start an organization *later*, so rounds of critical feedback and updating could occur *before* you lock in decisions about hiring, org structure, name, culture, etc.

But at least it happened *at all*; if OpenAI had just said "yeah, we're gonna do alignment research!" and left it there, the outcome probably would have been far worse.

Also, if organizations release obviously bad plans but are then unresponsive to counter-arguments, researchers can go work at the orgs with better plans and avoid the orgs with worse plans. This encourages groups to compete to have the seemingly-sanest plan, which strikes me as a better equilibrium than the current one.