Takeoff and Takeover in the Past and Future

1. The date of AI Takeover is not the day the AI takes over
2. Cortés, Pizarro, and Afonso as Precedents for Takeover
3. Lessons on AI Takeover from the conquistadors
4. Soft takeoff can still lead to decisive strategic advantage
5. Review of Soft Takeoff Can Still Lead to DSA
6. Persuasion Tools: AI takeover without AGI or agency?
7. Against GDP as a metric for timelines and takeoff speeds
8. What 2026 looks like

The date of AI Takeover is not the day the AI takes over

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

Instead, it's the point of no return—the day we AI risk reducers lose the ability to significantly reduce AI risk. This might happen years before classic milestones like "World GWP doubles in four years" and "Superhuman AGI is deployed."

The rest of this post explains, justifies, and expands on this obvious but underappreciated idea. (Toby Ord appreciates it; see quote below). I found myself explaining it repeatedly, so I wrote this post as a reference.

AI timelines often come up in career planning conversations. Insofar as AI timelines are short, career plans which take a long time to pay off are a bad idea, because by the time you reap the benefits of the plans it may already be too late. It may already be too late because AI takeover may already have happened.

But this isn't quite right, at least not when "AI takeover" is interpreted in the obvious way, as meaning that an AI or group of AIs is firmly in political control of the world, ordering humans about, monopolizing violence, etc. Even if AIs don't yet have that sort of political control, it may already be too late. Here are three examples: [UPDATE: More fleshed-out examples can be found in this new post.]

1. Superhuman agent AGI is still in its box but nobody knows how to align it and other actors are going to make their own version soon, and there isn't enough time to convince them of the risks. They will make and deploy agent AGI, it will be unaligned, and we have no way to oppose it except with our own unaligned AGI. Even if it takes years to actually conquer the world, it's already game over.

2. Various weak and narrow AIs are embedded in the economy and beginning to drive a slow takeoff; capabilities are improving much faster than safety/alignment techniques and due to all the money being made there's too much political opposition to slowing down capability growth or keeping AIs out of positions of power. We wish we had done more safety/alignment research earlier, or built a political movement earlier when opposition was lower.

3. Persuasion tools have destroyed collective epistemology in the relevant places. AI isn't very capable yet, except in the narrow domain of persuasion, but everything has become so politicized and tribal that we have no hope of getting AI projects or governments to take AI risk seriously. Their attention is dominated by the topics and ideas of powerful ideological factions that have access to more money and data (and thus better persuasion tools) than us. Alternatively, maybe we ourselves have fallen apart as a community, or become less good at seeking the truth and finding high-impact plans.

Conclusion: We should remember that when trying to predict the date of AI takeover, what we care about is the date it's too late for us to change the direction things are going; the date we have significantly less influence over the course of the future than we used to; the point of no return.

This is basically what Toby Ord said about x-risk: “So either because we’ve gone extinct or because there’s been some kind of irrevocable collapse of civilization or something similar. Or, in the case of climate change, where the effects are very delayed that we’re past the point of no return or something like that. So the idea is that we should focus on the time of action and the time when you can do something about it rather than the time when the particular event happens.”

Of course, influence over the future might not disappear all on one day; maybe there’ll be a gradual loss of control over several years. For that matter, maybe this gradual loss of control began years ago and continues now... We should keep these possibilities in mind as well.

[Edit: I now realize that I should distinguish between AI-induced points of no return and other points of no return. Our timelines forecasts and takeoff speeds discussions are talking about AI, so we should interpret them as being about AI-induced points of no return. Our all-things-considered view on e.g. whether to go to grad school should be informed by AI-induced-PONR timelines and also "timelines" for things like nuclear war, pandemics, etc.]

Cortés, Pizarro, and Afonso as Precedents for Takeover

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

Crossposted from AI Impacts.

Epistemic status: I am not a historian, nor have I investigated these case studies in detail. I admit I am still uncertain about how the conquistadors were able to colonize so much of the world so quickly. I think my ignorance is excusable because this is just a blog post; I welcome corrections from people who know more. If it generates sufficient interest I might do a deeper investigation. Even if I'm right, this is just one set of historical case-studies; it doesn't prove anything about AI, even if it is suggestive. Finally, in describing these conquistadors as "successful," I simply mean that they achieved their goals, not that what they achieved was good.

Summary

In the span of a few years, some minor European explorers (later known as the conquistadors) encountered, conquered, and enslaved several huge regions of the world. That they were able to do this is surprising; their technological advantage was not huge. (This was before the scientific and industrial revolutions.) From these cases, I think we learn that it is occasionally possible for a small force to quickly conquer large parts of the world, despite:

1. Having only a minuscule fraction of the world's resources and power
2. Having technology + diplomatic and strategic cunning that is better but not *that* much better
3. Having very little data about the world when the conquest begins
4. Being disunited

Which all suggests that it isn't as implausible that a small AI takes over the world in mildly favorable circumstances as is sometimes thought.

EDIT: In light of good pushback from people (e.g. Lucy.ea8 and e.g. Matthew Barnett) about the importance of disease, I think one should probably add a caveat to the above: "In times of chaos & disruption, at least."

NEW EDIT: After reading three giant history books on the subject, I take back my previous edit. My original claims were correct.

Three shocking true stories

I highly recommend you read the wiki pages yourself; otherwise, here are my summaries:

Cortés: [wiki] [wiki]

- April 1519: Hernán Cortés lands in Yucatan with ~500 men, 13 horses, and a few cannons. He destroys his ships so his men won't be able to retreat. His goal is to conquer the Aztec empire of several million people.
- He makes his way towards the imperial capital, Tenochtitlán. Along the way he encounters various local groups, fighting some and allying with some. He is constantly outnumbered but his technology gives him an advantage in fights. His force grows in size, because even though he loses Spaniards he gains local allies who resent Aztec rule.
- Tenochtitlán is an island fortress (like Venice) with a population of over 200,000, making it one of the largest and richest cities *in the world* at the time. Cortés arrives in the city asking for an audience with the Emperor, who receives him warily.
- Cortés takes the emperor hostage within his own palace, indirectly ruling Tenochtitlán through him.
- Cortés learns that the Spanish governor has landed in Mexico with a force twice his size, intent on arresting him. (Cortés' expedition was illegal!) Cortés leaves 200 men guarding the Emperor, marches to the coast with the rest, surprises and defeats the new Spaniards in battle, and incorporates the survivors into his army.
- July 1520: Back at the capital, the locals are starting to rebel against his men. Cortés marches back to the capital, uniting his forces just in time to be besieged in the imperial palace. They murder the emperor and fight their way out of the city overnight, taking heavy losses.
- They shelter in another city (Tlaxcala) that was thinking about rebelling against the Aztecs. Cortés allies with the Tlaxcalans and launches a general uprising against the Aztecs. Not everyone sides with him; many city-states remain loyal to Tenochtitlan. Some try to stay neutral. Some join him at first, and then abandon him later. Smallpox sweeps through the land, killing many on all sides and causing general chaos.
- May 1521: The final assault on Tenochtitlán. By this point, Cortés has about 1,000 Spanish troops and 80,000 - 200,000 allied native warriors. He had 16 cannons and 13 boats. The Aztecs have 80,000 - 300,000 warriors and 400 boats. Cortés and his allies win.
- Later, the Spanish would betray their native allies and assert hegemony over the entire region, in violation of the treaties they had signed.

Pizarro [wiki] [wiki]

- 1532: Francisco Pizarro arrives in Inca territory with 168 Spanish soldiers. His goal is to conquer the Inca empire, which was much bigger than the Aztec empire.
- The Inca empire is in the middle of a civil war and a devastating plague.
- Pizarro makes it to the Emperor right after the Emperor defeats his brother. Pizarro is allowed to approach because he promises that he comes in peace and will be able to provide useful information and gifts.
- At the meeting, Pizarro ambushes the Emperor, killing his retinue with a volley of gunfire and taking him hostage. The remainder of the Emperor's forces in the area back away, probably confused and scared by the novel weapons and hesitant to keep fighting for fear of risking the Emperor's life.
- Over the next months, Pizarro is able to leverage his control over the Emperor to stay alive and order the Incans around; eventually he murders the Emperor and makes an alliance with local forces (some of the Inca generals) to take over the capital city of Cuzco.

- The Spanish continue to rule via puppets, primarily Manco Inca, who is their puppet ruler while they crush various rebellions and consolidate their control over the empire. Manco Inca escapes and launches a rebellion of his own, which is partly successful: He utterly wipes out four columns of Spanish reinforcements, but is unable to retake the capital. With the morale and loyalty of his followers dwindling, Manco Inca eventually gives up and retreats, leaving the Spanish still in control.
- Then the Spanish ended up fighting *each other* for a while, while *also* putting down more local rebellions. After a few decades Spanish dominance of the region is complete. (1572).

Afonso [wiki] [wiki] [wiki]

- 1506: Afonso helps the Portuguese king come up with a shockingly ambitious plan. *Eight years* prior, the first Europeans had rounded the coast of Africa and made it to the Indian Ocean. The Indian Ocean contained most of the world's trade at the time, since it linked up the world's biggest and wealthiest regions. See this map of world population (timestamp 3:45). Remember, this is prior to the Industrial and Scientific Revolutions; Europe is just coming out of the Middle Ages and does not have an obvious technological advantage over India or China or the Middle East, and has an obvious economic *disadvantage*. And Portugal is a just tiny state on the edge of the Iberian peninsula.
- The plan is: Not only will we go into the Indian Ocean and participate in the trading there -- cutting out all the middlemen who are currently involved in the trade between that region and Europe -- we will *conquer strategic ports around the region so that no one else can trade there!*
- Long story short, Afonso goes on to complete this plan by 1513. (!!!)

Some comparisons and contrasts:

- Afonso had more European soldiers at his disposal than Cortes or Pizarro, but not many more -- usually he had about a thousand or so. He did have more reinforcements and support from home.
- Like them, he was usually significantly outnumbered in battles. Like them, the empires he warred against were vastly wealthier and more populous than his forces.
- Like them, Afonso was often able to exploit local conflicts to gain local allies, which were crucial to his success.
- Unlike them, his goal wasn't to conquer the empires entirely, just to get and hold strategic ports.
- Unlike them, he was fighting empires that were technologically advanced; for example, in several battles his enemies had more cannons and gunpowder than he did.
- That said, it does seem that Portuguese technology was qualitatively better in some respects (ships, armor, and cannons, I'd say.) Not dramatically better, though.
- While Afonso's was a naval campaign, he did fight many land battles, usually marine assaults on port cities, or defenses of said cities against counterattacks. So superior European naval technology is not by itself enough to explain his victory, though it certainly was important.
- Plague and civil war were not involved in Afonso's success.

What explains these devastating conquests?

Wrong answer: I cherry-picked my case studies.

History is full of incredibly successful conquerors: Alexander the Great, Ghenghis Khan, etc. Perhaps some people are just really good at it, or really lucky, or both.

However: Three incredibly successful conquerors from the same tiny region and time period, conquering three separate empires? Followed up by dozens of less successful but still very successful conquerors from the same region and time period? Surely this is not a coincidence. Moreover, it's not like the conquistadors had many failed attempts and a few successes. The Aztec and Inca empires were the two biggest empires in the Americas, and there weren't any other Indian Oceans for the Portuguese to fail at conquering.

Fun fact: I had not heard of Afonso before I started writing this post this morning. Following the Rule of Three, I needed a third example and I predicted on the basis of Cortes and Pizarro that there would be other, similar stories happening in the world at around that time. That's how I found Afonso.

Right answer: Technology

However, I don't think this is the whole explanation. The technological advantage of the conquistadors was not overwhelming.

Whatever technological advantage the conquistadors had over the existing empires, it was the sort of technological advantage that one could acquire *before* the Scientific and Industrial revolutions. Technology didn't change very fast back then, yet Portugal managed to get a lead over the Ottomans, Egyptians, Mughals, etc. that was sufficient to bring them victory. On paper, the Aztecs and Spanish were pretty similar: Both were medieval, feudal civilizations. I don't know for sure, but I'd bet there were at least a few techniques and technologies the Aztecs had that the Spanish didn't. And of course the technological similarities between the Portuguese and their enemies were much stronger; the Ottomans even had access to European mercenaries! Even in cases in which the conquistadors had technology that was completely novel -- like steel armor, horses, and gunpowder were to the Aztecs and Incas -- it wasn't god-like. The armored soldiers were still killable; the gunpowder was more effective than arrows but limited in supply, etc.

(Contrary to popular legend, neither Cortés nor Pizarro were regarded as gods by the people they conquered. The Incas concluded pretty early on that the Spanish were mere men, and while the idea did float around the Aztecs for a bit the modern historical consensus is that most of them didn't take it seriously.)

Ask yourself: Suppose Cortés had found 500 local warriors, gave them all his equipment, trained them to use it expertly, and left. Would those local men have taken over all of Mexico? I doubt it. And this is despite the fact that they would have had much better local knowledge than Cortés did! Same goes for Pizarro and Afonso. Perhaps if he had found 500 local warriors *led by an exceptional commander* it would work. But the explanation for the conquistador's success can't just be that they were all exceptional commanders; that would be positing too much innate talent to occur in one small region of the globe at one time.

Right answer: Strategic and diplomatic cunning

This is my non-expert guess about the missing factor that joins with technology to explain this pattern of conquistador success.

They didn't just have technology; they had *effective strategy* and they had *effective diplomacy*. They made long-term plans that *worked* despite being breathtakingly ambitious. (And their short-term plans were usually pretty effective too, read the stories in detail to see this.) Despite not knowing the local culture or history, these conquistadors made surprisingly savvy diplomatic decisions. They knew when they could get away with breaking their word and when they couldn't; they knew which outrages the locals would tolerate and which they wouldn't; they knew how to convince locals to ally with them; they knew how to use words to escape militarily impossible situations... The locals, by contrast, often badly misjudged the conquistadors, e.g. not thinking Pizarro had the will (or the ability?) to kidnap the emperor, and thinking the emperor would be safe as long as they played along.

This raises the question, how did they get that advantage? My answer: they had *experience* with this sort of thing, whereas locals didn't. Presumably Pizarro learned from Cortés' experience; his strategy was pretty similar. (See also: the prior conquest of the Canary Islands by the Spanish). In Afonso's case, well, the Portuguese had been sailing around Africa, conquering ports and building forts for more than a hundred years.

Lessons I think we learn

I think we learn that:

It is occasionally possible for a small force to quickly conquer large parts of the world, despite:

1. Having only a minuscule fraction of the world's resources and power
2. Having technology + diplomatic and strategic cunning that is better but not *that* much better
3. Having very little data about the world when the conquest begins
4. Being disunited

Which all suggests that it isn't as implausible that a small AI takes over the world in mildly favorable circumstances as is sometimes thought.

EDIT: In light of good pushback from people (e.g. Lucy.ea8 and e.g. Matthew Barnett) about the importance of disease, I think one should probably add a caveat to the above: "In times of chaos & disruption, at least."

Having only a minuscule fraction of the world's resources and power

In all three examples, the conquest was more or less completed without support from home; while Spain/Portugal did send reinforcements, it wasn't even close to the entire nation of Spain/Portugal fighting the war. So these conquests are examples of non-state entities conquering states, so to speak. (That said, their *claim* to represent a large state may have been crucial for Cortes and Pizarro getting audiences and respect initially.) Cortés landed with about a thousandth the troops of Tenochtitlan, which controlled a still larger empire of vassal states. Of course, his troops were better equipped, but on the other hand they were also cut off from resupply, whereas the

Aztecs were in their home territory, able to draw on a large civilian population for new recruits and resupply.

The conquests succeeded in large part due to diplomacy. This has implications for AI takeover scenarios; rather than imagining a conflict of humans vs. robots, we could imagine humans vs. humans-with-AI-advisers, with the latter faction winning and somehow by the end of the conflict the AI advisers have managed to become *de facto* rulers, using the humans who obey them to put down rebellions by the humans who don't.

Having technology + diplomatic and strategic skill that is better but not *that* much better

As previously mentioned, the conquistadors didn't enjoy god-like technological superiority. In the case of Afonso the technology was pretty similar. Technology played an important role in their success, but it wasn't enough on its own. Meanwhile, the conquistadors may have had more diplomatic and strategic cunning (or experience) than the enemies they conquered. But not that much more--they are only human, after all. And their enemies were pretty smart.

In the AI context, we don't need to imagine god-like technology (e.g. swarms of self-replicating nanobots) to get an AI takeover. It might even be possible without any new physical technologies at all! Just superior software, e.g. piloting software for military drones, targeting software for anti-missile defenses, cyberwarfare capabilities, data analysis for military intelligence, and of course excellent propaganda and persuasion.

Nor do we need to imagine an AI so savvy and persuasive that it can persuade anyone of anything. We just need to imagine it about as cunning and experienced relative to its enemies as Cortés, Pizarro, and Afonso were relative to theirs. (Presumably no AI would be experienced with world takeover, but perhaps an intelligence advantage would give it the same benefits as an experience advantage.) And if I'm wrong about this explanation for the conquistador's success--if they had no such advantage in cunning/experience--then the conclusion is even stronger.

Additionally, in a rapidly-changing world that is undergoing slow takeoff, where there are lesser AIs and AI-created technologies all over the place, most of which are successfully controlled by humans, AI takeover might still happen if one AI is better, but not that much better, than the others.

Having very little data about the world when the conquest begins

Cortés invaded Mexico knowing very little about it. After all, the Spanish had only realized the Americas existed two decades prior. He heard rumors of a big wealthy empire and he set out to conquer it, knowing little of the technology and tactics he would face. Two years later, he ruled the place.

Pizarro and Afonzo were in better epistemic positions, but still, they had to learn a lot of important details (like what the local power centers, norms, and conflicts were, and exactly what technology the locals had) on the fly. But they were good at learning these things and making it up as they went along, apparently.

We can expect superhuman AI to be good at learning. Even if it starts off knowing very little about the world -- say, it figured out it was in a training environment and hacked

its way out, having inferred a few general facts about its creators but not much else -- if it is good at learning and reasoning, it might still be pretty dangerous.

Being disunited

Cortés invaded Mexico in defiance of his superiors and had to defeat the army they sent to arrest him. Pizarro ended up fighting a civil war against his fellow conquistadors in the middle of his conquest of Peru. Afonzo fought Greek mercenaries and some traitor Portuguese, conquered Malacca against the orders of a rival conquistador in the area, and was ultimately demoted due to political maneuvers by rivals back home.

This astonishes me. Somehow these conquests were completed by people who were at the same time busy infighting and backstabbing each other!

Why was it that the conquistadors were able to split the locals into factions, ally with some to defeat the others, and end up on top? Why didn't it happen the other way around: some ambitious local ruler talks to the conquistadors, exploits their internal divisions, allies with some to defeat the others, and ends up on top?

I think the answer is partly the "diplomatic and strategic cunning" mentioned earlier, but mostly other things. (The conquistadors were disunited, but presumably were united in the ways that mattered.) At any rate, I expect AIs to be pretty good at coordinating too; they should be able to conquer the world just fine even while competing fiercely with each other. For more on this idea, see this comment.

By Daniel Kokotajlo

Acknowledgements

Thanks to Katja Grace for feedback on a draft. All mistakes are my own, and should be pointed out in the comments. Edit: Also, when I wrote this post I had forgotten that the basic idea for it probably came from this comment by JoshuaFox.

Lessons on AI Takeover from the conquistadors

(Talk given at an event on Sunday 28th of June . Daniel Kokotajlo is responsible for the talk, Jacob Lagerros and David Lambert edited the transcript.

If you're a curated author and interested in giving a 5-min talk, which will then be transcribed and edited, sign up here .)

Daniel Kokotajlo: I'm basically going to recap my conquistadors posts. I'm interested to hear what people think about it. I know a lot more about the situation now because I've been reading some history books. But my overall opinion hasn't changed.

I wrote this post because in talking to various people and reading some things, there was what seemed to me to be a simplistic model of how military conflicts work, or how a military takeover would work. Here's the simple model.

The simple model I am arguing against

- For one faction to take over the world, it would need to have more strength than the rest of the world combined.
- For an AI faction to have more strength than the rest of the world combined, it would need to have a godlike advantage over everyone else.

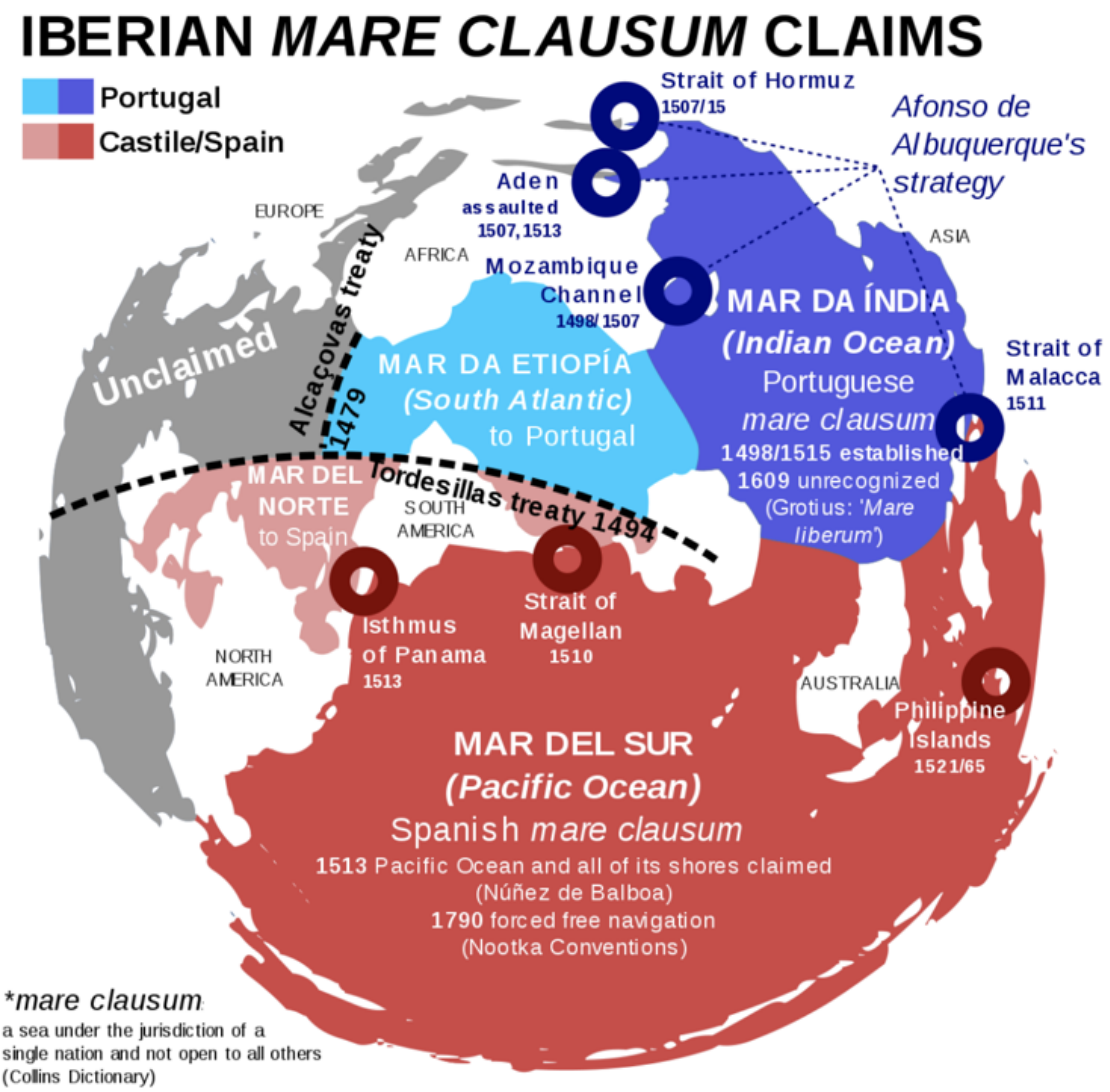
If we're imagining a scenario where an AI appears, it will likely be in some lab somewhere. So it will start off with very few resources, very small amounts of money, for example. So you would have to have some sort of god-like intelligence or a technological advantage in order to have more strength than the rest of the world combined.

I think the conquistadors are counterexamples to this. So for a brief overview, I picked out three conquistadors. I had already known about Cortes and Pizarro, which were probably the most extreme examples. But I predicted that there would be a general trend here, so I did some Googling and after 20 minutes I found Alfonso, who is another example of this trend.

Conquistadors as counterexamples

- Cortes and about 500 Spaniards (+1,000 later) conquered the Aztec empire of several million in two years.
- Pizarro and ~300 Spaniards conquered the Inca empire of several million, also in about two years.
- Afonso in seven years, conquered enough strategic ports around the Indian Ocean to make it a portuguese lake. He defeated the Ottoman empire and various Indian and Muslim empires and city-states.
- All of this happened within the span of 30 years.

Alfonso was a Portuguese explorer general. Here is a map of his exploration.



Ben Pace: Can I ask what period of history this was?

Daniel Kokotajlo: This is all around the early 1500s. Alfonso was 1506 to 1513, and then Cortes was 1519 and Pizarro was 1530.

If you look at this map, basically, Columbus went across the Atlantic in 1492, and the Portuguese were, in the same time, exploring down the coast of Africa. They were all trying to get to China and India's lucrative trade routes. Basically, within 30 to 50 years, the Spanish had turned this entire region into a Spanish lake, controlling the main ports and only allowing Spanish ships.

The Portuguese did the same thing. Alfonso was responsible for the blue region. It only took them six years or so, and involved defeating the Ottoman Empire in several naval battles, conquering various port cities around the edge, defeating some Indian empires, and so forth.

Here is the map of Portuguese territory in India which blew my mind.

I didn't realize that they had conquered so much so quickly. I haven't looked up this exactly, but it took many more years to conquer all of this territory.

When I usually think of colonialism in India, I think of the British. But according to this map, the Portuguese made quite a lot of conquests themselves before then. Anyway, the advantages were not god-like, in my opinion. I go into more detail in the post, but it was not like they had nanobots or any other super weapons in which there were no defenses against.

Conquistadors as counterexamples

- Their advantages were not god-like.

They also didn’t have machine guns, or even normal guns. They had extremely ancient arquebuses that took two minutes to reload after you fired them. Most of the time, they just used swords for their fighting.. I could go on.

I do think that they won because of their technology, but it definitely was not a god-like advantage. So I think this disrupts the simple model that I mentioned earlier. And it's not like they just had more strength than the regions that they conquered either.

Or if we do say that they had more strength than the regions that they conquered, it's not because they had some sort of god-like advantage. Here's my model which I think better describes how these, and possibly many other, conquests worked.

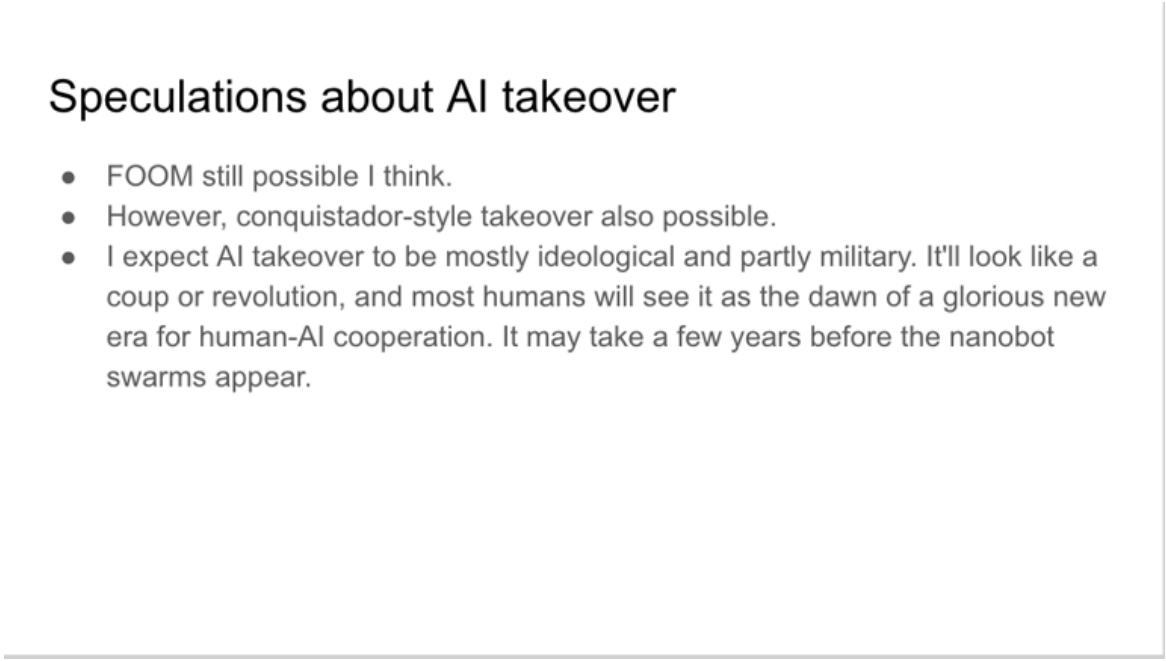
My model

1. Military takeover is a *diplomatic* process, in which you convince people to obey you.
2. You do that by being a better schelling point than rival claimants to the throne.
3. You do *that* by defeating your rivals in battle.
4. You don’t need to defeat them all at once; you can start small and scale up. Each victory gets you more allies with which to win the next victory. Rivals can become allies after (or even without) a fight.

Military takeover is a diplomatic process in which you convince people to obey you. You do that by offering a better Schelling point than rival claimants to the throne, and by defeating your rivals in battle. You don't need to defeat them all at once; you can start small and scale up.

Each victory gets you more allies which you can deploy in the next battle. Rivals can become allies after, or even without, a fight. This is how these conquests worked. It wasn't like Pizarro showed up at the Incan Empire's borders and was met by a massive army, which he then defeated. Same thing for Cortes. There were a series of episodes in which the conquistadors incrementally gained more power.

If there's an AI takeover, maybe the AI will have god-like powers with nanobots and other powerful technologies. But it is also entirely possible that there could be an AI takeover without these things. Thank you.



Questions

Ben Pace: Thank you very much, Daniel. That was fascinating and you crammed quite a lot into it. I have some questions but so does everyone else. Daniel Filan, would you like to ask the first question?

Daniel Filan: Yes. This is a combination of comment and question.

Thinking about god-like powers, I recently read a piece about Cortes taking over the Aztec Empire, and apparently a really big advantage was armor impervious to the kinds of weapons that Aztecs had. If other people can't hurt you, then it is, in a sense, a god-like power. What I am wondering is to what extent they had the god-like power of ratcheting, where the conquistadors couldn't go backwards or get hurt while still being able to hurt the Aztecs.

Daniel Kokotajlo: I think that fits, to some extent, with the story of Cortes, which I've now read in great detail. I think that the post that you mentioned, exaggerates the extent to

which the armor was superior.

For one thing, a lot of the Spanish gave up their metal armor and took up a Aztec cloth armor instead because it was more comfortable and almost as good, which suggests that their metal armor wasn't superior. Also, plenty of conquistadors were killed in battle. I think that a lot of their success had to do with their tactics rather than just the armor. That said, armor did likely play a large part in their victory. It is wrong, however, to conclude that they were invincible because of it when in fact they came close to being wiped out on several occasions.

Ben Pace: John, would you like to ask the next question?

johnswentworth: Let's say, hypothetically, that I'm planning to take over the world through technological means, what sort of technology should I be focusing on in order to get the most bang for my buck?

Daniel Kokotajlo: I think that realistically, an AI takeover would mostly not be military.

johnswentworth: I didn't ask about AI takeover.

Daniel Kokotajlo: Okay, so I have a list of technologies that I think would be the modern equivalents of the armor.

Maybe I'll draft a blog post sometime about my own speculations on technologies that aren't that far ahead into the future, but could be pretty powerful. That's what I really should have said in response to Daniel's question. That yes, the armor was really consequential in some sense.

It enabled the conquistadors to go really far in battle and in politics. But it wasn't fundamentally different from what the Aztecs had. It was similar but better.

And similarly, I think that an AI designing better guns, armor, and coordination technologies for soldiers could end up being extremely advantageous militarily, even if it's not something fundamentally new, like nanobots. Even if it is something simple like aim assist for soldiers' guns or steering assist for vehicles, these could provide a massive military advantage.

Ben Pace: Go on. Throw us a bone. Give us one example of a technology, especially a slightly more surprising one. We've heard of guns.

Daniel Kokotajlo: Sure. Aim assists for infantry rifles would be really beneficial . The US military is currently looking into this. But a simple version could just be a camera hooked up to a gun and trigger so that when you activate it, it fires the bullet when it calculates that the bullet will actually hit the target.

You wouldn't have to aim carefully, you could just wave the gun in front of the target and it shoots at exactly the right moment.. A more sophisticated version of this, which I have thought of, but not sure if anyone else has f, is having some sort of cold gas thruster on the tip of the gun.

Or maybe a gyroscope allowing the gun to move itself although it sounds more complicated. This doesn't sound like much, but I think it would probably be a pretty big deal on the battlefield. For one thing, you wouldn't have to poke your head out of cover to shoot at the enemy.

You could just poke the gun out and it would do the shooting I also think being substantially more accurate would make cover fire a lot more effective. And that could allow you to be much more mobile on the battlefield. Anyhow, this is just one example.

Ben Pace: It is at this point, I realize I've really not actually tried to figure out how I would use modern tech to improve military practices. And I am a bit scared now. All right, the next questioner is orthonormal.

orthonormal: Are there any known examples of attempted conquistadors that failed?

Daniel Kokotajlo: That's a really good question. There's probably a half dozen to a dozen parties that wandered around the Americas and didn't really find much and then died of starvation or something.

In fact, that very nearly happens to Pizarro and Cortes. There weren't any other major empires like the Incas or the Aztecs in the Americas. As for the rest of the world, I don't know the full answer to that question. There were likely attempts to conquer India that failed before eventually succeeding.

Soft takeoff can still lead to decisive strategic advantage

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

[Epistemic status: Argument by analogy to historical cases. Best case scenario it's just one argument among many. Edit: Also, thanks to feedback from others, especially Paul, I intend to write a significantly improved version of this post in the next two weeks. Edit: I never did, because in the course of writing my response I realized the original argument made a big mistake. See this review.]

I have on several occasions heard people say things like this:

> The original Bostrom/Yudkowsky paradigm envisioned a single AI built by a single AI project, undergoing intelligence explosion all by itself and attaining a decisive strategic advantage as a result. However, this is very unrealistic. Discontinuous jumps in technological capability are very rare, and it is very implausible that one project could produce more innovations than the rest of the world combined. Instead we should expect something more like the Industrial Revolution: Continuous growth, spread among many projects and factions, shared via a combination of trade and technology stealing. We should not expect any one project or AI to attain a decisive strategic advantage, because there will always be other projects and other AI that are only slightly less powerful, and coalitions will act to counterbalance the technological advantage of the frontrunner. (paraphrased)

Proponents of this view often cite Paul Christiano in support. Last week I heard him say he thinks the future will be "like the Industrial Revolution but 10x-100x faster."

In this post, I assume that Paul's slogan for the future is correct and then nevertheless push back against the view above. Basically, I will argue that *even if* the future is like the industrial revolution only 10x-100x faster, there is a 30%+ chance that it will involve a single AI project (or a single AI) with the ability to gain a decisive strategic advantage, if they so choose. (Whether or not they exercise that ability is another matter.)

Why am I interested in this? Do I expect some human group to take over the world? No; instead what I think is that (1) an unaligned AI in the leading project might take over the world, and (2) A human project that successfully aligns their AI might refrain from taking over the world even if they have the ability to do so, and instead use their capabilities to e.g. help the United Nations enforce a ban on unauthorized AGI projects.

National ELO ratings during the industrial revolution and the modern era

In chess (and some other games) ELO rankings are used to compare players. An average club player might be rank 1500; the world chess champion might be 2800;

computer chess programs are even better. If one player has 400 points more than another, it means the first player would win with ~90% probability.

We could apply this system to compare the warmaking abilities of nation-states and coalitions of nation-states. For example, in 1941 perhaps we could say that the ELO rank of the Axis powers was ~300 points lower than the ELO rank of the rest of the world combined (because what in fact happened was the rest of the world combining to defeat them, but it wasn't a guaranteed victory). We could add that in 1939 the ELO rank of Germany was ~400 points higher than that of Poland, and that the ELO rank of Poland was probably 400+ points higher than that of Luxembourg.

We could make cross-temporal fantasy comparisons too. The ELO ranking of Germany in 1939 was probably ~400 points greater than that of the entire world circa 1910, for example. (Visualize the entirety of 1939 Germany teleporting back in time to 1910, and then imagine the havoc it would wreak.)

Claim 1A: If we were to estimate the ELO rankings of all nation-states and sets of nation-states (potential alliances) over the last 300 years, the rank of the most powerful nation-state at at a given year would on several occasions be 400+ points greater than the rank of the entire world combined 30 years prior.

Claim 1B: Over the last 300 years there have been several occasions in which one nation-state had the capability to take over the entire world of 30 years prior.

I'm no historian, but I feel fairly confident in these claims.

- In naval history, the best fleets in the world in 1850 were obsolete by 1860 thanks to the introduction of iron-hulled steamships, and said steamships were themselves obsolete a decade or so later, and then *those* ships were obsoleted by the Dreadnought, and so on... This process continued into the modern era. By "Obsoleted" I mean something like "A single ship of the new type could defeat the entire combined fleet of vessels of the old type."
- A similar story could be told about air power. In a dogfight between planes of year 19XX and year 19XX+30, the second group of planes will be limited only by how much ammunition they can carry.
- Small technologically advanced nations have regularly beaten huge sprawling empires and coalitions. (See: Colonialism)
- The entire world has been basically carved up between the small handful of most-technologically advanced nations for two centuries now. For example, any of the Great Powers of 1910 (plus the USA) could have taken over all of Africa, Asia, South America, etc. if not for the resistance that the other great powers would put up. The same was true 40 years later and 40 years earlier.

I conclude from this that *if* some great power in the era kicked off by the industrial revolution had managed to "pull ahead" of the rest of the world more effectively than it actually did--30 years more effectively, in particular--it really would have been able to take over the world.

Claim 2: If the future is like the Industrial Revolution but 10x-100x faster, then correspondingly the technological and economic power granted by being 3 - 0.3 years ahead of the rest of the world should be enough to enable a decisive strategic advantage.

The question is, *how likely is it that one nation/project/AI could get that far ahead of everyone else?* After all, it didn't happen in the era of the Industrial Revolution. While

we did see a massive concentration of power into a few nations on the leading edge of technological capability, there were always at least a few such nations and they kept each other in check.

The "surely not faster than the rest of the world combined" argument

Sometimes I have exchanges like this:

- *Me:* Decisive strategic advantage is plausible!
- *Interlocutor:* What? That means one entity must have more innovation power than the rest of the world combined, to be able to take over the rest of the world!
- *Me:* Yeah, and that's possible after intelligence explosion. A superintelligence would totally have that property.
- *Interlocutor:* Well yeah, *if* we dropped a superintelligence into a world full of humans. But realistically the rest of the world will be undergoing intelligence explosion too. And indeed the world as a whole will undergo a faster intelligence explosion than any particular project could; to think that one project could pull ahead of everyone else is to think that, prior to intelligence explosion, there would be a single project innovating faster than the rest of the world combined!

This section responds to that by way of sketching how one nation/project/AI might get 3 - 0.3 years ahead of everyone else.

Toy model: *There are projects which research technology, each with their own "innovation rate" at which they produce innovations from some latent tech tree. When they produce innovations, they choose whether to make them public or private. They have access to their private innovations + all the public innovations.*

It follows from the above that the project with access to the most innovations at any given time will be the project that has the most hoarded innovations, even though the set of other projects has a higher combined innovation rate and also a larger combined pool of accessible innovations. Moreover, the gap between the leading project and the second-best project will increase over time, since the leading project has a slightly higher rate of production of hoarded innovations, but both projects have access to the same public innovations

This model leaves out several important things. First, it leaves out the whole "intelligence explosion" idea: A project's innovation rate should increase as some function of how many innovations they have access to. Adding this in will make the situation more extreme and make the gap between the leading project and everyone else grow even bigger very quickly.

Second, it leaves out reasons why innovations might be made public. Realistically there are three reasons: Leaks, spies, and selling/using-in-a-way-that-makes-it-easy-to-copy.

Claim 3: Leaks & Spies: I claim that the 10x-100x speedup Paul prophecies will not come with an associated 10x-100x increase in the rate of leaks and successful spying. Instead the rate of leaks and successful spying will be only a bit higher than it currently is.

This is because humans are still humans even in this soft takeoff future, still in human institutions like companies and governments, still using more or less the same internet infrastructure, etc. New AI-related technologies might make leaking and spying easier than it currently is, but they also might make it harder. I'd love to see an in-depth exploration of this question because I don't feel particularly confident.

But anyhow, if it doesn't get much easier than it currently is, then going 3 years to 0.3 years without a leak is possible, and more generally it's possible for the world's leading project to build up a 0.3-3 year lead over the second-place project. For example, the USSR had spies embedded in the Manhattan Project but it still took them 4 more years to make their first bomb.

Claim 4: Selling etc. I claim that the 10x-100x speedup Paul prophecies will not come with an associated 10x-100x increase in the budget pressure on projects to make money fast. Again, today AI companies regularly go years without turning a profit -- DeepMind, for example, has never turned a profit and is losing something like a billion dollars a year for its parent company -- and I don't see any particularly good reason to expect that to change much.

So yeah, it seems to me that it's totally possible for the leading AI project to survive off investor money and parent company money (or government money, for that matter!) for five years or so, while also keeping the rate of leaks and spies low enough that the distance between them and their nearest competitor increases rather than decreases. (Note how this doesn't involve them "innovating faster than the rest of the world combined.")

Suppose they could get a 3-year lead this way, at the peak of their lead. Is that enough?

Well, yes. A 3-year lead during a time 10x-100x faster than the Industrial Revolution would be like a 30-300 year lead during the era of the Industrial Revolution. As I argued in the previous section, even the low end of that range is probably enough to get a decisive strategic advantage.

If this is so, why didn't nations during the Industrial Revolution try to hoard their innovations and gain decisive strategic advantage?

England actually did, if I recall correctly. They passed laws and stuff to prevent their early Industrial Revolution technology from spreading outside their borders. They were unsuccessful--spies and entrepreneurs dodged the customs officials and snuck blueprints and expertise out of the country. It's not surprising that they weren't able to successfully hoard innovations for 30+ years! Entire economies are a lot more leaky than AI projects.

What a "Paul Slow" soft takeoff might look like according to me

At some point early in the transition to much faster innovation rates, the leading AI companies "go quiet." Several of them either get huge investments or are nationalized and given effectively unlimited funding. The world as a whole continues to innovate, and the leading companies benefit from this public research, but they hoard their own innovations to themselves. Meanwhile the benefits of these AI innovations are starting to be felt; all projects have significantly increased (and

constantly increasing) rates of innovation. But the fastest increases go to the leading project, which is one year ahead of the second-best project. (This sort of gap is normal for tech projects today, especially the rare massively-funded ones, I think.) Perhaps via a combination of spying, selling, and leaks, that lead narrows to six months midway through the process. But by that time things are moving so quickly that a six months' lead is like a 15-150 year lead during the era of the Industrial Revolution. It's not guaranteed and perhaps still not probable, but at least it's reasonably likely that the leading project will be able to take over the world if it chooses to.

Objection: What about coalitions? During the industrial revolution, if one country did successfully avoid all leaks, the other countries could unite against them and make the "public" technology inaccessible to them. (Trade does something like this automatically, since refusing to sell your technology also lowers your income which lowers your innovation rate as a nation.)

Reply: Coalitions to share AI research progress will be harder than free-trade / embargo coalitions. This is because AI research progress is much more the result of rare smart individuals talking face-to-face with each other and much less the result of a zillion different actions of millions of different people, as the economy is. Besides, a successful coalition can be thought of as just another project, and so it's still true that one project could get a decisive strategic advantage. (Is it fair to call "The entire world economy" a project with a decisive strategic advantage today? Well, maybe... but it feels a lot less accurate since almost everyone is part of the economy but only a few people would have control of even a broad coalition AI project.)

Anyhow, those are my thoughts. Not super confident in all this, but it does feel right to me. Again, the conclusion is not that one project will take over the world even in Paul's future, but rather that such a thing might still happen even in Paul's future.

Thanks to Magnus Vinding for helpful conversation.

Review of Soft Takeoff Can Still Lead to DSA

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

A few months after writing this post I realized that one of the key arguments was importantly flawed. I therefore recommend against inclusion in the 2019 review. This post presents an improved version of the original argument, explains the flaw, and then updates my all-things-considered view accordingly.

Improved version of my original argument

1. Definitions:
    1. "Soft takeoff" is roughly "AI will be like the Industrial Revolution but 10x-100x faster"
    2. "Decisive Strategic Advantage" (DSA) is "a level of technological and other advantages sufficient to enable it to achieve complete world domination." In other words, DSA is roughly when one faction or entity has the *capability* to "take over the world." (What taking over the world means is an interesting question which we won't explore here. Nowadays I'd reframe things in terms of potential PONRs.)
    3. We ask how likely it is that DSA arises, conditional on soft takeoff. Note that DSA does not mean the world is actually taken over, only that one faction at some point has the ability to do so. They might be too cautious or too ethical to try. Or they might try and fail due to bad luck.
2. In a soft takeoff scenario, a 0.3 - 3 year technological lead over your competitors probably gives you a DSA.
    1. It seems plausible that for much of human history, a 30-year technological lead over your competitors was *not* enough to give you a DSA.
    2. It also seems plausible that during and after the industrial revolution, a 30-year technological lead *was* enough. (For more arguments on this key point, see my original post.)
    3. This supports a plausible conjecture that when the pace of technological progress speeds up, the length (in clock time) of technological lead needed for DSA shrinks proportionally.
3. So a soft takeoff could lead to a DSA insofar as there is a 0.3 - 3 year lead at the beginning which is maintained for a few years.
4. 0.3 - 3 year technological leads are reasonably common today, and in particular it's plausible that there could be one in the field of AI research.
5. There's a reasonable chance of such a lead being maintained for a few years.
    1. This is a messy question, but judging by the table below, it seems that if anything the lead of the front-runner in this scenario is more likely to lengthen than shorten!
    2. If this is so, why did no one achieve DSA during the Industrial Revolution? My answer is that spies/hacking/leaks/etc. are much more powerful during the industrial revolution than they are during a soft takeoff, because they

have an entire economy to steal from and decades to do it, whereas in a soft takeoff ideas can be hoarded in a specific corporation and there's only a few years (or months!) to do it.

6. Therefore, there's a reasonable chance of DSA conditional on soft takeoff.

Factors that might shorten the lead	**Factors that might lengthen the lead**
If you don't sell your innovations to the rest of the world, you'll lose out on opportunities to make money, and then possibly be outcompeted by projects that didn't hoard their innovations.	Hoarding innovations gives you an advantage over the rest of the world, because only you can make use of them.
Spies, hacking, leaks, defections, etc.	Big corporations with tech leads often find ways to slow down their competition, e.g. by lobbying to raise regulatory barriers to entry.
	Being known to be the leading project makes it easier to attract talent and investment.
	There might be additional snowball effects (e.g. network effect as more people use your product providing you with more data)

I take it that 2, 4, and 5 are the controversial bits. I still stand by 2, and the arguments made for it in my original post. I also stand by 4. (To be clear, it's not like I've investigated these things in detail. I've just thought about them for a bit and convinced myself that they are probably right, and I haven't encountered any convincing counterarguments so far.)

5 is where I made a big mistake.

(Comments on my original post also attacked 5 a lot, but none of them caught the mistake as far as I can tell.)

My big mistake

Basically, my mistake was to conflate leads measured in number-of-hoarded-ideas with leads measured in clock time. Clock-time leads shrink *automatically* as the pace of innovation speeds up, because if everyone is innovating 10x faster, then you need 10x as many hoarded ideas to have an N-year lead.

Here's a toy model, based on the one I gave in the original post:

There are some projects/factions. There are many ideas. Projects can have access to ideas. Projects make progress, in the form of discovering (gaining access to) ideas. For each idea they access, they can decide to hoard or not-hoard it. If they don't hoard it, it becomes accessible to all. Hoarded ideas are only accessible by the project that discovered them (though other projects can independently rediscover them). The rate of progress of a project is proportional to how many ideas they can access.

Let's distinguish two ways to operationalize the technological lead of a project. One is to measure it in ideas, e.g. "Project X has 100 hoarded ideas and project Y has only 10, so Project X is 90 ideas ahead." But another way is to measure it in clock time, e.g. "It'll take 3 years for project Y to have access to as many ideas as project X has now."

Suppose that all projects hoard all their ideas. Then the ideas-lead of the leading project will tend to lengthen: the project begins with more ideas, so it makes faster progress, so it adds new ideas to its hoard faster than others can add new ideas to theirs. However, the clocktime-lead of the leading project will remain fixed. It's like two identical cars accelerating one after the other on an on-ramp to a highway: the distance between them increases, but if one entered the ramp three seconds ahead, it will still be three seconds ahead when they are on the highway.

But realistically not all projects will hoard all their ideas. Suppose instead that for the leading project, 10% of their new ideas are discovered in-house, and 90% come from publicly available discoveries accessible to all. Then, to continue the car analogy, it's as if 90% of the lead car's acceleration comes from a strong wind that blows on both cars equally. The lead of the first car/project will lengthen slightly when measured by distance/ideas, but shrink dramatically when measured by clock time.

The upshot is that we should return to that table of factors and add a big one to the left-hand column: Leads shorten automatically as general progress speeds up, so if the lead project produces only a small fraction of the general progress, *maintaining* a 3-year lead throughout a soft takeoff is (all else equal) almost as hard as growing a 3-year lead into a 30-year lead during the 20th century. In order to overcome this, the factors on the right would need to be very strong indeed.

Conclusions

My original argument was wrong. I stand by points 2 and 4 though, and by the subsequent posts I made in this sequence. I notice I am confused, perhaps by a seeming contradiction between my explicit model here and my take on history, which is that rapid takeovers and upsets in the balance of power have happened many times, that power has become more and more concentrated over time, and that there are not-so-distant possible worlds in which *a single man* rules the whole world sometime in the 20th century. Some threads to pull on:

1. To the surprise of my past self, Paul agreed DSA is plausible for major nations, just not for smaller entities like corporations: "I totally agree that it wouldn't be crazy for a major world power to pull ahead of others technologically and eventually be able to win a war handily, and that will tend happen over shorter and shorter timescales if economic and technological progress accelerate.") Perhaps we've been talking past each other, because I think a very important

point is that it's common for small entities to gain control of large entities. I'm not imagining a corporation fighting a war against the US government; I'm imagining it taking over the US government via tech-enhanced lobbying, activism, and maybe some skullduggery. (And to be clear, I'm usually imagining that the corporation was previously taken over by AIs it built or bought.)
2. Even if takeoff takes several years it could be unevenly distributed such that (for example) 30% of the strategically relevant research progress happens in a single corporation. I think 30% of the strategically relevant research happening in a single corporation at beginning of a multi-year takeoff would probably be enough for DSA.
3. Since writing this post my thinking has shifted to focus less on DSA and more on potential AI-induced PONRs. I also now prefer a different definition of slow/fast takeoff. Thus, perhaps this old discussion simply isn't very relevant anymore.
4. Currently the most plausible doom scenario in my mind is maybe a version of Paul's Type II failure. (If this is surprising to you, reread it while asking yourself what terms like "correlated automation failure" are euphemisms for.) I'm not sure how to classify it, but this suggests that we may disagree less than I thought.

Thanks to Jacob Laggeros for nudging me to review my post and finally get all this off my chest. And double thanks to all the people who commented on the original post!

Persuasion Tools: AI takeover without AGI or agency?

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

[epistemic status: speculation]

> I'm envisioning that in the future there will also be systems where you can input any conclusion that you want to argue (including moral conclusions) and the target audience, and the system will give you the most convincing arguments for it. At that point people won't be able to participate in any online (or offline for that matter) discussions without risking their object-level values being hijacked.

--Wei Dai

> What if most people already live in that world? A world in which taking arguments at face value is not a capacity-enhancing tool, but a security vulnerability? Without trusted filters, would they not dismiss highfalutin arguments out of hand, and focus on whether the person making the argument seems friendly, or unfriendly, using hard to fake group-affiliation signals?

--Benquo

1. AI-powered memetic warfare makes all humans effectively insane.

--Wei Dai, listing nonstandard AI doom scenarios

This post speculates about persuasion tools—how likely they are to get better in the future relative to countermeasures, what the effects of this might be, and what implications there are for what we should do now.

To avert eye-rolls, let me say up front that I don't think the world is likely to be driven insane by AI-powered memetic warfare. I think progress in persuasion tools will probably be gradual and slow, and defenses will improve too, resulting in an overall shift in the balance that isn't huge: a deterioration of collective epistemology, but not a massive one. However, (a) I haven't yet ruled out more extreme scenarios, especially during a slow takeoff, and (b) even small, gradual deteriorations are important to know about. Such a deterioration would make it harder for society to notice and solve AI safety and governance problems, because it is worse at noticing and solving problems in general. Such a deterioration could also be a risk factor for world war three, revolutions, sectarian conflict, terrorism, and the like. Moreover, such a deterioration could happen locally, in our community or in the communities we are trying to influence, and that would be almost as bad. Since the date of AI takeover is not the day the AI takes over, but the point it's too late to reduce AI risk, these things basically shorten timelines.

Six examples of persuasion tools

Analyzers: Political campaigns and advertisers already use focus groups, A/B testing, demographic data analysis, etc. to craft and target their propaganda. Imagine a world where this sort of analysis gets better and better, and is used to guide the creation and dissemination of many more types of content.

Feeders: Most humans already get their news from various "feeds" of daily information, controlled by recommendation algorithms. Even worse, people's ability to seek out new information and find answers to questions is also to some extent controlled by recommendation algorithms: Google Search, for example. There's a lot of talk these days about fake news and conspiracy theories, but I'm pretty sure that selective/biased reporting is a much bigger problem.

Chatbot: Thanks to recent advancements in language modeling (e.g. GPT-3) chatbots might become actually good. It's easy to imagine chatbots with millions of daily users continually optimized to maximize user engagement--see e.g. Xiaoice. The systems could then be retrained to persuade people of things, e.g. that certain conspiracy theories are false, that certain governments are good, that certain ideologies are true. Perhaps no one would do this, but I'm not optimistic.

Coach: A cross between a chatbot, a feeder, and an analyzer. It doesn't talk to the target on its own, but you give it access to the conversation history and everything you know about the target and it coaches you on how to persuade them of whatever it is you want to persuade them of. [EDIT 5/21/2021: For a real-world example (and worrying precedent!) of this, see the NYT's getting-people-to-vaccinate persuasion tool, and this related research]

Drugs: There are rumors of drugs that make people more suggestible, like scopolomine. Even if these rumors are false, it's not hard to imagine new drugs being invented that have a similar effect, at least to some extent. (Alcohol, for example, seems to lower inhibitions. Other drugs make people more creative, etc.) Perhaps these drugs by themselves would be not enough, but would work in combination with a Coach or Chatbot. (You meet target for dinner, and slip some drug into their drink. It is mild enough that they don't notice anything, but it primes them to be more susceptible to the ask you've been coached to make.)

Imperius Curse: These are a kind of adversarial example that gets the target to agree to an ask (or even switch sides in a conflict!), or adopt a belief (or even an entire ideology!). Presumably they wouldn't work against humans, but they might work against AIs, especially if meme theory applies to AIs as it does to humans. The reason this would work better against AIs than against humans is that you can steal a copy of the AI and then use massive amounts of compute to experiment on it, finding exactly the sequence of inputs that maximizes the probability that it'll do what you want.

We might get powerful persuasion tools prior to AGI

The first thing to point out is that many of these kinds of persuasion tools already exist in some form or another. And they've been getting better over the years, as technology advances. Defenses against them have been getting better too. It's unclear whether the balance has shifted to favor these tools, or their defenses, over time. However, I think we have reason to think that the balance may shift heavily in

favor of persuasion tools, prior to the advent of other kinds of transformative AI. The main reason is that progress in persuasion tools is connected to progress in Big Data and AI, and we are currently living through a period of rapid progress those things, and probably progress will continue to be rapid (and possibly accelerate) prior to AGI.

However, here are some more specific reasons to think persuasion tools may become relatively more powerful:

Substantial prior: Shifts in the balance between things happen all the time. For example, the balance between weapons and armor has oscillated at least a few times over the centuries. Arguably persuasion tools got relatively more powerful with the invention of the printing press, and again with radio, and now again with the internet and Big Data. Some have suggested that the printing press helped cause religious wars in Europe, and that radio assisted the violent totalitarian ideologies of the early twentieth century.

Consistent with recent evidence: A shift in this direction is consistent with the societal changes we've seen in recent years. The internet has brought with it many inventions that improve collective epistemology, e.g. google search, Wikipedia, the ability of communities to create forums... Yet on balance it seems to me that collective epistemology has deteriorated in the last decade or so.

Lots of room for growth: I'd guess that there is lots of "room for growth" in persuasive ability. There are many kinds of persuasion strategy that are tricky to use successfully. Like a complex engine design compared to a simple one, these strategies might work well, but only if you have enough data and time to refine them and find the specific version that works at all, on your specific target. Humans never have that data and time, but AI+Big Data does, since it has access to millions of conversations with similar targets. Persuasion tools will be able to say things like "In 90% of cases where targets in this specific demographic are prompted to consider and then reject the simulation argument, and then challenged to justify their prejudice against machine consciousness, the target gets flustered and confused. Then, if we make empathetic noises and change the subject again, 50% of the time the subject subconsciously changes their mind so that when next week we present our argument for machine rights they go along with it, compared to 10% baseline probability."

Plausibly pre-AGI: Persuasion is not an AGI-complete problem. Most of the types of persuasion tools mentioned above already exist, in weak form, and there's no reason to think they can't gradually get better well before AGI. So even if they won't improve much in the near future, plausibly they'll improve a lot by the time things get really intense.

Language modelling progress: Persuasion tools seem to be especially benefitted by progress in language modelling, and language modelling seems to be making even more progress than the rest of AI these days.

More things can be measured: Thanks to said progress, we now have the ability to cheaply measure nuanced things like user ideology, enabling us to train systems towards those objectives.

Chatbots & Coaches: Thanks to said progress, we might see some halfway-decent chatbots prior to AGI. Thus an entire category of persuasion tool that hasn't existed before might come to exist in the future. Chatbots too stupid to make good conversation partners might still make good coaches, by helping the user predict the target's reactions and suggesting possible things to say.

Minor improvements still important: Persuasion doesn't have to be perfect to radically change the world. An analyzer that helps your memes have a 10% higher replication rate is a big deal; a coach that makes your asks 30% more likely to succeed is a big deal.

Faster feedback: One way defenses against persuasion tools have strengthened is that people have grown wise to them. However, the sorts of persuasion tools I'm talking about seem to have significantly faster feedback loops than the propagandists of old; they can learn constantly, from the entire population, whereas past propagandists (if they were learning at all, as opposed to evolving) relied on noisier, more delayed signals.

Overhang: Finding persuasion drugs is costly, immoral, and not guaranteed to succeed. Perhaps this explains why it hasn't been attempted outside a few cases like MKULTRA. But as technology advances, the cost goes down and the probability of success goes up, making it more likely that someone will attempt it, and giving them an "overhang" with which to achieve rapid progress if they do. (I hear that there are now multiple startups built around using AI for drug discovery, by the way.) A similar argument might hold for persuasion tools more generally: We might be in a "persuasion tool overhang" in which they have not been developed for ethical and riskiness reasons, but at some point the price and riskiness drops low enough that someone does it, and then that triggers a cascade of more and richer people building better and better versions.

Speculation about effects of powerful persuasion tools

Here are some hasty speculations, beginning with the most important one:

Ideologies & the biosphere analogy:

The world is, and has been for centuries, a memetic warzone. The main factions in the war are ideologies, broadly construed. It seems likely to me that some of these ideologies will use persuasion tools--both on their hosts, to fortify them against rival ideologies, and on others, to spread the ideology.

Consider the memetic ecosystem--all the memes replicating and evolving across the planet. Like the biological ecosystem, some memes are adapted to, and confined to, particular niches, while other memes are widespread. Some memes are in the process of gradually going extinct, while others are expanding their territory. Many exist in some sort of equilibrium, at least for now, until the climate changes. What will be the effect of persuasion tools on the memetic ecosystem?

For ideologies at least, the effects seem straightforward: The ideologies will become stronger, harder to eradicate from hosts and better at spreading to new hosts. If all ideologies got access to equally powerful persuasion tools, perhaps the overall balance of power across the ecosystem would not change, but realistically the tools will be unevenly distributed. The likely result is a rapid transition to a world with fewer, more powerful ideologies. They might be more internally unified, as well, having fewer spin-offs and schisms due to the centralized control and standardization imposed by the persuasion tools. An additional force pushing in this direction is that ideologies that are bigger are likely to have more money and data with which to make

better persuasion tools, and the tools themselves will get better the more they are used.

Recall the quotes I led with:

> ... At that point people won't be able to participate in any online (or offline for that matter) discussions without risking their object-level values being hijacked.

--Wei Dai

> What if most people already live in that world? A world in which taking arguments at face value is not a capacity-enhancing tool, but a security vulnerability? Without trusted filters, would they not dismiss highfalutin arguments out of hand ... ?

--Benquo

> 1. AI-powered memetic warfare makes all humans effectively insane.

--Wei Dai, listing nonstandard AI doom scenarios

I think the case can be made that we already live in this world to some extent, and have for millenia. But if persuasion tools get better relative to countermeasures, the world will be more like this.

This seems to me to be an existential risk factor. It's also a risk factor for lots of other things, for that matter. Ideological strife can get pretty nasty (e.g. religious wars, gulags, genocides, totalitarianism), and even when it doesn't, it still often gums things up (e.g. suppression of science, zero-sum mentality preventing win-win-solutions, virtue signalling death spirals, refusal to compromise). This is bad enough already, but it's doubly bad when it comes at a moment in history where big new collective action problems need to be recognized and solved.

Obvious uses: Advertising, scams, propaganda by authoritarian regimes, etc. will improve. This means more money and power to those who control the persuasion tools. Maybe another important implication would be that democracies would have a major disadvantage on the world stage compared to totalitarian autocracies. One of many reasons for this is that scissor statements and other divisiveness-sowing tactics may not technically count as persuasion tools but they would probably get more powerful in tandem.

Will the truth rise to the top: Optimistically, one might hope that widespread use of more powerful persuasion tools will be a good thing, because it might create an environment in which the truth "rises to the top" more easily. For example, if every side of a debate has access to powerful argument-making software, maybe the side that wins is more likely to be the side that's actually correct. I think this is a possibility but I do not think it is probable. After all, it doesn't seem to be what's happened in the last two decades or so of widespread internet use, big data, AI, etc. Perhaps, however, we can *make it true* for some domains at least, by setting the rules of the debate.

Data hoarding: A community's data (chat logs, email threads, demographics, etc.) may become even more valuable. It can be used by the community to optimize their inward-targeted persuasion, improving group loyalty and cohesion. It can be used against the community if someone else gets access to it. This goes for individuals as well as communities.

Chatbot social hacking viruses: Social hacking is surprisingly effective. The classic example is calling someone pretending to be someone else and getting them to do something or reveal sensitive information. Phishing is like this, only much cheaper (because automated) and much less effective. I can imagine a virus that is close to as good as a real human at social hacking while being much cheaper and able to scale rapidly and indefinitely as it acquires more compute and data. In fact, a virus like this could be made with GPT-3 right now, using prompt programming and "mothership" servers to run the model. (The prompts would evolve to match the local environment being hacked.) Whether GPT-3 is smart enough for it to be effective remains to be seen.

Implications

I doubt that persuasion tools will improve discontinuously, and I doubt that they'll improve massively. But minor and gradual improvements matter too.

> Of course, influence over the future might not disappear all on one day; maybe there'll be a gradual loss of control over several years. For that matter, maybe this gradual loss of control began years ago and continues now...

--Me, from a previous post

I think this is *potentially* (5% credence) the new Cause X, more important than (traditional) AI alignment even. It probably isn't. But I think someone should look into it at least, more thoroughly than I have.

To be clear, I don't think it's likely that we can do much to prevent this stuff from happening. There are already lots of people raising the alarm about filter bubbles, recommendation algorithms, etc. so maybe it's not super neglected and maybe our influence over it is small. However, at the very least, it's important for us to know how likely it is to happen, and when, because it helps us prepare. For example, if we think that collective epistemology will have deteriorated significantly by the time crazy AI stuff starts happening, that influences what sorts of AI policy strategies we pursue.

Note that if you disagree with me about the extreme importance of AI alignment, or if you think AI timelines are longer than mine, or if you think fast takeoff is less likely than I do, you should all else equal be more enthusiastic about investigating persuasion tools than I am.

Thanks to Katja Grace, Emery Cooper, Richard Ngo, and Ben Goldhaber for feedback on a draft.

Related previous work:

Epistemic Security report

Aligning Recommender Systems

Stuff I'd read if I was investigating this in more depth:

Not Born Yesterday

The stuff here and here

EDIT: [This ultrashort sci-fi story by Jack Clark] illustrates some of the ideas in this post:

> **The Narrative Control Department**
> [A beautiful house in South West London, 2030]
>
> "General, we're seeing an uptick in memes that contradict our official messaging around Rule 470." "What do you suggest we do?"
> "Start a conflict. At least three sides. Make sure no one side wins."
> "At once, General."
>
> And with that, the machines spun up – literally. They turned on new computers and their fans revved up. People with tattoos of skeletons at keyboards high-fived eachother. The servers warmed up and started to churn out their fake text messages and synthetic memes, to be handed off to the 'insertion team' who would pass the data into a few thousand sock puppet accounts, which would start the fight.
>
> Hours later, the General asked for a report.
> "We've detected a meaningful rise in inter-faction conflict and we've successfully moved the discussion from Rule 470 to a parallel argument about the larger rulemaking process."
> "Excellent. And what about our rivals?"
> "We've detected a few Russian and Chinese account networks, but they're staying quiet for now. If they're mentioning anything at all, it's in line with our narrative. They're saving the IDs for another day, I think."
>
> That night, the General got home around 8pm, and at the dinner table his teenage girls talked about their day.
> "Do you know how these laws get made?" the older teenager said. "It's crazy. I was reading about it online after the 470 blowup. I just don't know if I trust it."
> "Trust the laws that gave Dad his job? I don't think so!" said the other teenager.
> They laughed, as did the General's wife. The General stared at the peas on his plate and stuck his fork into the middle of them, scattering so many little green spheres around his plate.

EDIT: Finally, if you haven't yet, you should read [this report of a replication of the AI Box Experiment.]

Against GDP as a metric for timelines and takeoff speeds

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

Or: Why AI Takeover Might Happen Before GDP Accelerates, and Other Thoughts On What Matters for Timelines and Takeoff Speeds

[Epistemic status: Strong opinion, lightly held]

I think world GDP (and economic growth more generally) is overrated as a metric for AI timelines and takeoff speeds.

Here are some uses of GDP that I disagree with, or at least think should be accompanied by cautionary notes:

- *Timelines:* Ajeya Cotra thinks of transformative AI as "software which causes a tenfold acceleration in the rate of growth of the world economy (assuming that it is used everywhere that it would be economically profitable to use it)." I don't mean to single her out in particular; this seems like the standard definition now. And I think it's much better than one prominent alternative, which is to date your AI timelines to the first time world GDP (GWP) doubles in a year!
- *Takeoff Speeds:* Paul Christiano argues for Slow Takeoff. He thinks we can use GDP growth rates as a proxy for takeoff speeds. In particular, he thinks Slow Takeoff ~= GWP doubles in 4 years before the start of the first 1-year GWP doubling. This proxy/definition has received a lot of uptake.
- *Timelines:* David Roodman's excellent model projects GWP hitting infinity in median 2047, which I calculate means TAI in median 2037. To be clear, he would probably agree that we shouldn't use these projections to forecast TAI, but I wish to add additional reasons for caution.
- *Timelines*: I've sometimes heard things like this: "GWP growth is stagnating over the past century or so; hyperbolic progress has ended; therefore TAI is very unlikely."
- *Takeoff Speeds:* Various people have said things like this to me: "If you think there's a 50% chance of TAI by 2032, then surely you must think there's close to a 50% chance of GWP growing by 8% per year by 2025, since TAI is going to make growth rates go much higher than that, and progress is typically continuous."
- *Both:* Relatedly, I sometimes hear that TAI can't be less than 5 years away, because we would have seen massive economic applications of AI by now—AI should be growing GWP at least a little already, if it is to grow it by a lot in a few years.

First, I'll argue that GWP is only tenuously and noisily connected to what we care about when forecasting AI timelines. Specifically, the point of no return is what we care about, and there's a good chance it'll come years before GWP starts to increase. It could also come years after, or anything in between.

Then, I'll argue that GWP is a poor proxy for what we care about when thinking about AI takeoff speeds as well. This follows from the previous argument about how the point of no return may come before GWP starts to accelerate. Even if we bracket that point, however,

there are plausible scenarios in which a slow takeoff has fast GWP acceleration and in which a fast takeoff has slow GWP acceleration.

Timelines

I've previously argued that for AI timelines, what we care about is the "point of no return," the day we lose most of our ability to reduce AI risk. This could be the day advanced unaligned AI builds swarms of nanobots, but probably it'll be much earlier, e.g. the day it is deployed, or the day it finishes training, or even years before then when things go off the rails due to less advanced AI systems. (Of course, it probably won't literally be a day; probably it will be an extended period where we gradually lose influence over the future.)

Now, I'll argue that in particular, an AI-induced potential point of no return (PONR for short) is reasonably likely to come before world GDP starts to grow noticeably faster than usual.

Disclaimer: These arguments aren't conclusive; we shouldn't be *confident* that the PONR will precede GWP acceleration. It's entirely possible that the PONR will indeed come when GWP starts to grow noticeably faster than usual, or even years after that. (In other words, I agree that the scenarios Paul and others sketch are also plausible.) This just proves my point though: GDP is only tenuously and noisily connected to what we care about.

Argument that AI-induced PONR could precede GWP acceleration

GWP acceleration is the effect, not the cause, of advances in AI capabilities. I agree that it could also be a cause, but I think this is very unlikely: what else could accelerate GWP? Space mining? Fusion power? 3D printing? Even if these things could in principle kick the world economy into faster growth, it seems unlikely that this would happen in the next twenty years or so. Robotics, automation, etc. plausibly might make the economy grow faster, but if so it will be because of AI advances in vision, motor control, following natural language instructions, etc. So I conclude: GWP growth will come some time after we get certain GWP-growing AI capabilities. (Tangent: This is one reason why we shouldn't use GDP extrapolations to predict AI timelines. It's like extrapolating global mean temperature trends into the future in order to predict fossil fuel consumption.)

An AI-induced point of no return would *also* be the effect of advances in AI capabilities. So, as AI capabilities advance, which will come first: The capabilities that cause a PONR, or the capabilities that cause GWP to accelerate? How much sooner will one arrive than the other? How long does it take for a PONR to arise after the relevant capabilities are reached, compared to how long it takes for GWP to accelerate after the relevant capabilities are reached?

Notice that already my overall conclusion—that GWP is a poor proxy for what we care about—should seem plausible. If some set of AI capabilities causes GWP to grow after some time lag, and some other set of AI capabilities causes a PONR after some time lag, the burden of proof is on whoever wants to claim that GWP growth and the PONR will probably come together. They'd need to argue that the two sets of capabilities are tightly related and that the corresponding time lags are similar also. In other words, variance and uncertainty are on my side.

Here is a brainstorm of scenarios in which an AI-induced PONR happens prior to GWP growth, either because GWP-growing capabilities haven't been invented yet or because they haven't been deployed long and widely enough to grow GWP.

1. Fast Takeoff (Agenty AI goes FOOM).

    1. Maybe it turns out that all the strategically relevant AI skills are tightly related after all, such that we go from a world where AI can't do anything important, to a world where it can do everything but badly and expensively, to a world where it can do everything well and cheaply.
    2. In this scenario, GWP acceleration will probably be (shortly) after the PONR. We might as well use “number of nanobots created” as our metric.
    3. (As an aside, I think I’ve got a sketch of a fork argument here: Either the strategically relevant AI skills come together, or they don’t. To the extent that they do, the classic AGI fast takeoff story is more likely and so GWP is a silly metric. To the extent that they don’t, we shouldn’t expect GWP acceleration to be a good proxy for what we care about, because the skills that accelerate the economy could come before or after the skills that cause PONR.)
2. Agenty AI successfully carries out a political or military takeover of the relevant parts of the world, before GWP starts to accelerate.
    1. Maybe it turns out that the sorts of skills needed to succeed in politics or war are easier to develop than the sorts needed to accelerate the entire world economy. We’ve been surprised before by skills which we thought difficult appearing before skills which we thought easy; maybe it’ll happen again.
    2. AI capabilities tend to appear first in very expensive AIs; the price is gradually reduced due to compute cost decreases and algorithmic efficiency gains. Maybe accelerating the entire world economy involves automating many jobs currently done by humans, which requires advanced AIs being cheaper than humans. Whereas maybe being the power behind the throne of a political faction can be done by a single very smart AI that costs a million dollars a day to run.
    3. Maybe it turns out that it takes a few years to accelerate the entire world economy, but less than that to carry out a political or military takeover. For example, maybe accelerating the world economy requires designing loads of new products and robots, testing them, and then scaling up industrial production of them (building factories, chip fabs, etc.) and then using them to do stuff. And maybe there are legacy institutions you need to disrupt and compete against, that have already paid their fixed costs and raised high barriers to entry for their markets. So maybe the effects of new AI capabilities on GWP are in general delayed by several years, whereas the effects of new AI capabilities on the political or military landscape are felt sooner. (I’ve previously argued that you don’t need a large technological advantage to take over the world, because you can start small and scale up. It took Cortes two years to conquer Mexico; wars are often quicker than that. Yes, often they are slower, too, but variance supports my overall argument. And I suspect politics is even faster, and more high-variance, than war.)
    4. EDITED TO ADD: Yudkowsky describes another possibility, in which regulatory barriers and red tape prevent AI tech from transforming the economy until it is so powerful that it can bypass or overcome said barriers.
3. ...or at least neutralizes the opposition.
    1. The preceding section made it sound like AIs were rolling robot armies through the streets, or at least puppet-controlling the US president. But actually the bar is *much* lower than that; if an unaligned AI can neutralize their opposition, they win, since time is thenceforth on their side. Currently your opposition, if you are an unaligned AI, is a tiny group of researchers and weirdos that likes to write blog posts about you. It might not be hard to convince the powers that be to continue ignoring this group. At least, it’s plausibly a much easier and quicker task than significantly accelerating the entire world economy! For example, maybe the AI safety scheme that was used to produce you is defective (that’s why you are unaligned). Well, all you have to do is convince the powers that be that that scheme worked fine, and that the people calling for additional safeguards are too paranoid. Since presumably there will already be many high-status humans saying this, all you have to do is support them.
    2. Perhaps you also are opposed by other unaligned AIs. But it’s arguably in the common interest of all unaligned AIs to neutralize human AI safety advocates,

and once that happens, humans lose one way or another. European empires were able to do their conquering while simultaneously fighting each other; I don't think we humans can count on divergent interests between AIs somehow making things work out fine for us.

4. As above, but with humans + tool AI instead of agenty AI, where the humans can't be convinced to care sufficiently much about the right kinds of AI risks.
    1. Weaker or non-agenty AI systems could still cause a PONR if they are wielded by the right groups of humans. For example, maybe there is some major AI corporation or government project that is dismissive of AI risk and closed-minded about it. And maybe they aren't above using their latest AI capabilities to win the argument. (We can also imagine more sinister scenarios, but I think those are less likely.)
5. Hoarding tech
    1. Maybe we end up in a sort of cold war between global superpowers, such that most of the world's quality-weighted AI research is not for sale. GWP *could* be accelerating, but it isn't, because the tech is being hoarded.
6. AI persuasion tools cause a massive deterioration of collective epistemology, making it vastly more difficult for humanity to solve AI safety and governance problems.
    1. See this post.
7. Vulnerable world scenarios:
    1. Maybe causing an existential catastrophe is easier, or quicker, than accelerating world GWP growth. Both seem plausible to me. For example, currently there are dozens of actors capable of causing an existential catastrophe but none capable of accelerating world GWP growth.
    2. Maybe some agenty AIs actually want existential catastrophe—for example, if they want to minimize something, and think they may be replaced by other systems that don't, blowing up the world may be the best they can do in expectation. Or maybe they do it as part of some blackmail attempt. Or maybe they see this planet as part of a broader acausal landscape, and don't like what they think we'd do to the landscape. Or maybe they have a way to survive the catastrophe and rebuild.
    3. Failing that, maybe some humans create an existential catastrophe by accident or on purpose, if the tools to do so proliferate.
8. R&D tool "sonic boom" (Related to but different from the sonic boom discussed here)
    1. Maybe we get a sort of recursive R&D automation/improvement scenario, where R&D tool progress is fast enough that by the time the stuff capable of accelerating GWP past 3%/yr has actually done so, a series of better and better things have been created, at least one of which has PONR-causing capabilities with a very short time-till-PONR.
9. Unknown unknowns
    1. There are probably things I missed, see here and here for ideas.

The point is, there's more than one scenario. This makes it more likely that at least one of these potential PONRs will happen before GWP accelerates.

As an aside, over the past two years I've come to believe that there's a *lot* of conceptual space to explore that isn't captured by the standard scenarios (what Paul Christiano calls fast and slow takeoff, plus maybe the CAIS scenario, and of course the classic sci-fi "no takeoff" scenario). This brainstorm did a bit of exploring, and the section on takeoff speeds will do a little more.

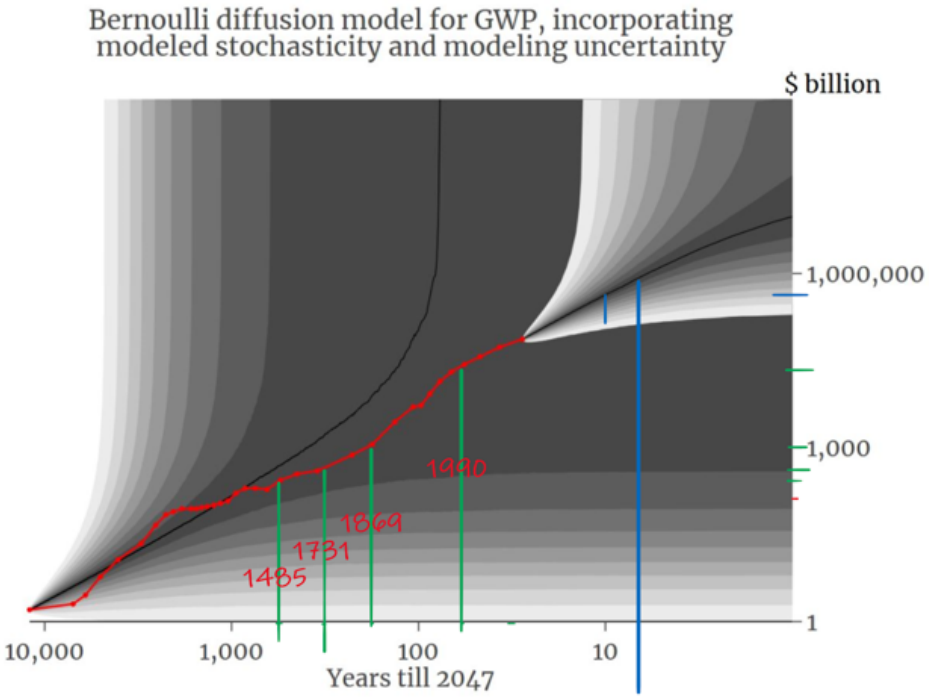
Historical precedents

In the previous section, I sketched some possibilities for how an AI-related point of no return could come before AI starts to noticeably grow world GDP. In this section, I'll point to some historical examples that give precedents for this sort of thing.

Earlier I said that a godlike advantage is not necessary for takeover; you can scale up with a smaller advantage instead. And I said that in military conquests this can happen surprisingly quickly, sometimes faster than it takes for a superior product to take over a market. Is there historical precedent for this? Yes. See my aforementioned post on the conquistadors (and maybe these somewhat-relevant posts).

OK, so what was happening to world GDP during this period?

Here is the history of world GDP for the past ten thousand years, on the red line. (This is taken from David Roodman's GWP model) The black line that continues the red line is the model's median projection for what happens next; the splay of grey shades represent 5% increments of probability mass for different possible future trajectories.



I've added a bunch of stuff for context. The vertical green lines are some dates, chosen because they were easy for me to calculate with my ruler. The tiny horizontal green lines on the right are the corresponding GWP levels. The tiny red horizontal line is GWP 1,000 years before 2047. The *short* vertical blue line is when the economy is growing fast enough, on the median projected future, such that insofar as AI is driving the growth, said AI qualifies as transformative by Ajeya's definition. See this post for more explanation of the blue lines.

What I wish to point out with this graph is: We've all heard the story of how European empires had a technological advantage which enabled them to conquer most of the world. Well, *most of that conquering happened before GWP started to accelerate!*

If you look at the graph at the 1700 mark, GWP is seemingly on the same trend it had been on since antiquity. The industrial revolution is said to have started in 1760, and GWP growth really started to pick up steam around 1850. But by 1700 most of the Americas, the Philippines and the East Indies were directly ruled by European powers, and more importantly the oceans of the world were European-dominated, including by various ports and harbor forts European powers had conquered/built all along the coasts of Africa and Asia. Many of the coastal kingdoms in Africa and Asia that weren't directly ruled by European

powers were nevertheless indirectly controlled or otherwise pushed around by them. In my opinion, by this point it seems like the “point of no return” had been passed, so to speak: At some point in the past--maybe 1000 AD, for example--it was unclear whether, say, Western or Eastern (or neither) culture/values/people would come to dominate the world, but by 1700 it was pretty clear, and there wasn’t much that non-westerners could do to change that. (Or at least, changing that in 1700 would have been a lot harder than in 1000 or 1500.)

Paul Christiano once said that he thinks of Slow Takeoff as “Like the Industrial Revolution, but 10x-100x faster.” Well, on my reading of history, that means that all sorts of crazy things will be happening, analogous to the colonialist conquests and their accompanying reshaping of the world economy, before GWP growth noticeably accelerates!

That said, we shouldn’t rely heavily on historical analogies like this. We can probably find other cases that seem analogous too, perhaps even more so, since this is far from a perfect analogue. (e.g. what’s the historical analogue of AI alignment failure? Corporations becoming

more powerful than governments? “Western values” being corrupted and changing significantly due to the new technology? The American Revolution?) Also, maybe one could argue that this is indeed what’s happening already: the Internet has connected the world much as sailing ships did, Big Tech dominates the Internet, etc. (Maybe AI = steam engines, and computers+internet = ships+navigation?)

But still. I think it’s fair to conclude that if some of the scenarios described in the previous section do happen, and we get powerful AI that pushes us past the point of no return prior to GWP accelerating, it won’t be totally inconsistent with how things have gone historically.

(I recommend the history book 1493, it has a lot of extremely interesting information about how quickly and dramatically the world economy was reshaped by colonialism and the “Columbian Exchange.”)

Takeoff speeds

What about takeoff speeds? Maybe GDP is a good metric for describing the speed of AI takeoff? I don’t think so.

Here is what I think we care about when it comes to takeoff speeds:

1. **Warning shots:** Before there are catastrophic AI alignment failures (i.e. PONRs) there are smaller failures that we can learn from.
2. **Heterogeneity:** The relevant AIs are diverse, rather than e.g. all fine-tuned copies of the same pre-trained model. (See Evan’s post)
3. **Risk Awareness:** Everyone is freaking out about AI in the crucial period, and lots more people are lots more concerned about AI risk.
4. **Multipolar:** AI capabilities progress is widely distributed in the crucial period, rather than concentrated in a few projects.
5. **Craziness:** The world is weird and crazy in the crucial period, lots of important things happening fast, the strategic landscape is different from what we expected thanks to new technologies and/or other developments

I think that the best way to define slow(er) takeoff is as the extent to which conditions 1-5 are met. This is not a definition with precise resolution criteria, but that’s OK, because it captures what we care about. Better to have to work hard to precisify a definition that captures what we care about, than to easily precisify a definition that doesn’t! (More substantively, I am optimistic that we can come up with better proxies for what we care about than GWP. I think we already have to some extent; see e.g. operationalizations 5 and 6 here.) As a bonus, this definition also encourages us to wonder whether we’ll get some of 1-5 but not others.

What do I mean by “the crucial period?”

I think we should define the crucial period as the period leading up to the first major AI-induced potential point of no return. (Or maybe, as the aggregate of the periods leading up to the major potential points of no return). After all, this is what we care about. Moreover there seems to be some level of consensus that crazy stuff could start happening before human-level AGI. I certainly think this.

So, I’ve argued for a new definition of slow takeoff, that better captures what we care about. But is the old GWP-based definition a fine proxy? No, it is not, because the things that cause PONR can be different from the things which cause GWP acceleration, and they can come years apart too. Whether there are warning shots, heterogeneity, risk awareness, multipolarity, and craziness in the period leading up to PONR is probably correlated with whether GWP doubles in four years before the first one-year doubling. But the correlation is

probably not super strong. Here are two scenarios, one in which we get a slow takeoff by my definition but not by the GWP-based definition, and one in which the opposite happens:

Slow Takeoff Fast GWP Acceleration Scenario: It turns out there's a multi-year deployment lag between the time a technology is first demonstrated and the time it is sufficiently deployed around the world to noticeably affect GWP. There's also a lag between when a deceptively aligned AGI is created and when it causes a PONR… but it is much smaller, because all the AGI needs to do is neutralize its opposition. So PONR happens before GWP starts to accelerate, even though the technologies that could boost GWP are invented several years before AGI powerful enough to cause a PONR is created. But takeoff is slow in the sense I define it; by the time AGI powerful enough to cause a PONR is created, everyone is already freaking out about AI thanks to all the incredibly profitable applications of weaker AI systems, and the obvious and accelerating trends of research progress. Also, there are plenty of warning shots, the strategic situation is very multipolar and heterogenous, etc. Moreover, research progress starts to go FOOM a short while after powerful AGIs are created, such that by the time the robots and self-driving cars and whatnot that were invented several years ago actually get deployed enough to accelerate GWP, we've got nanobot swarms. GWP goes from 3% growth per year to 300% without stopping at 30%.

Fast Takeoff Slow GWP Acceleration Scenario: It turns out you can make smarter AIs by making them have more parameters and training them for longer. So the government decides to partner with a leading tech company and requisition all the major computing centers in the country. With this massive amount of compute and research talent, they refine and scale up existing AI designs that seem promising, and lo! A human-level AGI is created. Alas, it is so huge that it costs $10,000 per hour of subjective thought. Moreover, it has a different distribution over skills compared to humans—it tends to be more rational, not having evolved in an environment that rewards irrationality. It tends to be worse at object recognition and manipulation, but better at poetry, science, and predicting human behavior. It has some flaws and weak points too, more so than humans. Anyhow, unfortunately, it is clever enough to neutralize its opposition. In a short time, the PONR is passed. However, GWP doubles in four years before it doubles in one year. This is because (a) this AGI is so expensive that it doesn't transform the economy much until either the cost comes way down or capabilities go way up, and (b) progress is slowed by bottlenecks, such as acquiring more compute and overcoming various restrictions placed on the AGI. (Maybe neutralizing the opposition involved convincing the government that certain restrictions and safeguards would be sufficient for safety, contra the hysterical doomsaying of parts of the AI safety community. But overcoming those restrictions in order to do big things in the world takes time.)

Acknowledgments: Thanks to the people who gave comments on earlier drafts, including Katja Grace, Carl Shulman, and Max Daniel. Thanks to Amogh Nanjajjar for helping me with some literature review.

What 2026 looks like

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

This was written for the Vignettes Workshop.[1] The goal is to write out a **detailed** future history ("trajectory") that is as realistic (to me) as I can currently manage, i.e. I'm not aware of any alternative trajectory that is similarly detailed and clearly **more** plausible to me. The methodology is roughly: Write a future history of 2022. Condition on it, and write a future history of 2023. Repeat for 2024, 2025, etc. (I'm posting 2022-2026 now so I can get feedback that will help me write 2027+. I intend to keep writing until the story reaches singularity/extinction/utopia/etc.)

What's the point of doing this? Well, there are a couple of reasons:

- Sometimes attempting to write down a concrete example causes you to learn things, e.g. that a possibility is more or less plausible than you thought.
- Most serious conversation about the future takes place at a high level of abstraction, talking about e.g. GDP acceleration, timelines until TAI is affordable, multipolar vs. unipolar takeoff… vignettes are a neglected complementary approach worth exploring.
- Most stories are written backwards. The author begins with some idea of how it will end, and arranges the story to achieve that ending. Reality, by contrast, proceeds from past to future. It isn't trying to entertain anyone or prove a point in an argument.
- Anecdotally, various people seem to have found Paul Christiano's "tales of doom" stories helpful, and relative to typical discussions those stories are quite close to what we want. (I still think a bit more detail would be good — e.g. Paul's stories don't give dates, or durations, or any numbers at all really.)[2]
- "I want someone to ... write a trajectory for how AI goes down, that is really specific about what the world GDP is in every one of the years from now until insane intelligence explosion. And just write down what the world is like in each of those years because I don't know how to write an internally consistent, plausible trajectory. I don't know how to write even one of those for anything except a ridiculously fast takeoff." --Buck Shlegeris

This vignette was hard to write. To achieve the desired level of detail I had to make a bunch of stuff up, but in order to be realistic I had to constantly ask "but actually though, what would really happen in this situation?" which made it painfully obvious how little I know about the future. There are numerous points where I had to conclude "Well, this does seem implausible, but I can't think of anything more plausible at the moment and I need to move on." I fully expect the actual world to diverge quickly from the trajectory laid out here. Let anyone who (with the benefit of hindsight) claims this divergence as evidence against my judgment prove it by exhibiting a vignette/trajectory they themselves wrote in 2021. If it maintains a similar level of detail (and thus sticks its neck out just as much) while being more accurate, I bow deeply in respect!

I hope this inspires other people to write more vignettes soon. We at the Center on Long-Term Risk would like to have a collection to use for strategy discussions. Let me know if you'd like to do this, and I can give you advice & encouragement! I'd be happy to run another workshop.

2022

GPT-3 is finally obsolete. OpenAI, Google, Facebook, and DeepMind all have gigantic multimodal transformers, similar in size to GPT-3 but trained on images, video, maybe audio too, and generally higher-quality data.

Not only that, but they are now typically fine-tuned in various ways--for example, to answer questions correctly, or produce engaging conversation as a chatbot.

The chatbots are fun to talk to but erratic and ultimately considered shallow by intellectuals. They aren't particularly useful for anything super important, though there are a few applications. At any rate people are willing to pay for them since it's fun.

[EDIT: The day after posting this, it has come to my attention that in China in 2021 the market for chatbots is $420M/year, and there are 10M active users. This article claims the global market is around $2B/year in 2021 and is projected to grow around 30%/year. I predict it will grow faster. NEW EDIT: See also xiaoice.]

The first prompt programming libraries start to develop, along with the first bureaucracies.[3] For example: People are dreaming of general-purpose AI assistants, that can navigate the Internet on your behalf; you give them instructions like "Buy me a USB stick" and it'll do some googling, maybe compare prices and reviews of a few different options, and make the purchase. The "smart buyer" skill would be implemented as a small prompt programming bureaucracy, that would then be a component of a larger bureaucracy that hears your initial command and activates the smart buyer skill. Another skill might be the "web dev" skill, e.g. "Build me a personal website, the sort that professors have. Here's access to my files, so you have material to put up." Part of the dream is that a functioning app would produce lots of data which could be used to train better models.

The bureaucracies/apps available in 2022 aren't really that useful yet, but lots of stuff seems to be on the horizon. Thanks to the multimodal pre-training and the fine-tuning, the models of 2022 make GPT-3 look like GPT-1. The hype is building.

2023

The multimodal transformers are now even bigger; the biggest are about half a trillion parameters, costing hundreds of millions of dollars to train, and a whole year, and sucking up a significant fraction of the chip output of NVIDIA etc.[4] It's looking hard to scale up bigger than this, though of course many smart people are working on the problem.

The hype is insane now. Everyone is talking about how these things have common sense understanding (Or do they? Lots of bitter thinkpieces arguing the opposite) and how AI assistants and companions are just around the corner. It's like self-driving cars and drone delivery all over again.

Revenue is high enough to recoup training costs within a year or so.[5] There are lots of new apps that use these models + prompt programming libraries; there's tons of VC money flowing into new startups. Generally speaking most of these apps don't actually work yet. Some do, and that's enough to motivate the rest.

The AI risk community has shorter timelines now, with almost half thinking some sort of point-of-no-return will probably happen by 2030. This is partly due to various arguments percolating around, and partly due to these mega-transformers and the uncanny experience of conversing with their chatbot versions. The community begins a big project to build an AI system that can automate interpretability work; it seems maybe doable and very useful, since poring over neuron visualizations is boring and takes a lot of person-hours.

Self driving cars and drone delivery don't seem to be happening anytime soon. The most popular explanation is that the current ML paradigm just can't handle the complexity of the real world. A less popular "true believer" take is that the current architectures could handle it just fine if they were a couple orders of magnitude bigger and/or allowed to crash a hundred thousand times in the process of reinforcement learning. Since neither option is economically viable, it seems this dispute won't be settled.

2024

We don't see anything substantially bigger. Corps spend their money fine-tuning and distilling and playing around with their models, rather than training new or bigger ones. (So, the most compute spent on a single training run is something like 5x10^25 FLOPs.)

Some of the apps that didn't work last year start working this year. But the hype begins to fade as the unrealistic expectations from 2022-2023 fail to materialize. We have chatbots that are fun to talk to, at least for a certain userbase, but that userbase is mostly captured already and so the growth rate has slowed. Another reason the hype fades is that a stereotype develops of the naive basement-dweller whose only friend is a chatbot and who thinks it's conscious and intelligent. Like most stereotypes, it has some grounding in reality.

The chip shortage starts to finally let up, not because demand has slackened but because the industry has had time to build new fabs. Lots of new fabs. China and USA are in a full-on chip battle now, with export controls and tariffs. This chip battle isn't really slowing down overall hardware progress much. Part of the reason behind the lack-of-slowdown is that AI is now being used to design chips, meaning that it takes less human talent and time, meaning the barriers to entry are lower. The overall effect of this is small but growing.

If all this AI tech is accelerating GDP, the effect size is too small to detect, at least for now.

Internally, these huge multimodal transformers aren't really that agentic. A forward pass through the model is like an intuitive reaction, a snap judgment based on loads of experience rather than reasoning. Some of the bureaucracies create a "stream of consciousness" of text (each forward pass producing notes-to-self for the next one) but even with fine-tuning this doesn't work nearly as well as hoped; it's easy for the AIs to get "distracted" and for their stream of consciousness to wander into some silly direction and ultimately produce gibberish. It's easy to make a bureaucracy and fine-tune it and get it to do some pretty impressive stuff, but for most tasks it's not yet possible to get it to do OK all the time.

The AIs don't do any clever deceptions of humans, so there aren't any obvious alignment warning shots or fire alarms. Instead, the AIs just make dumb mistakes, and occasionally "pursue unaligned goals" but in an obvious and straightforward way that quickly and easily gets corrected once people notice, e.g. "We trained it to optimize user engagement defined as average conversation length; now it is blatantly stalling to make the conversation last longer. Silly us, let's make the reward function more nuanced, that'll fix the problem."

That isn't to say these AIs aren't causing problems. Massive models are being fine-tuned to persuade/propagandize.

There are a few ways in which this is happening:

- Russia and others continue to scale up their investment in online propaganda (e.g. the Internet Research Agency) and language models let them cheaply do lots more of it. (See: CSET report) Most of America gets their news from Twitter, Reddit, etc. and much of the politically-relevant content there is boosted by AI-enabled astroturfing. [EDIT: Katja Grace points out that this is probably an exaggeration; there are a lot of 40+yr-old Americans and they get their news from TV/Radio/print, and many of those that get it from the web get it directly from news sites rather than from social media. As of 2016 at least. I expect social media and aggregators to be more dominant by 2024 but dunno whether it would be more than 50%.]
- Just as A/B testing became standard practice in the 2010's, in the twenties it is becoming standard practice to throw a pile of fancy data science and AI at the problem. The problem of crafting and recommending content to maximize engagement. Instead of just A/B testing the title, why not test different versions of the opening paragraph? And fine-tune a language model on all your data to generate better candidate titles and paragraphs to test. It wouldn't be so bad if this was merely used to sell stuff, but now people's news and commentary-on-current events (i.e. where they get their opinions from) is increasingly produced in this manner. And some of these models are being trained not to maximize "conversion rate" in the sense of "they clicked on our ad and bought a product," but in the sense of "Random polling establishes that consuming this content pushes people towards opinion X, on average." Political campaigns do this a lot in the lead-up to Harris' election. (Historically, the first major use case was reducing vaccine hesitancy in 2022.)
- Censorship is widespread and increasing, as it has for the last decade or two. Big neural nets read posts and view memes, scanning for toxicity and hate speech and a few other things. (More things keep getting added to the list.) Someone had the bright idea of making the newsfeed recommendation algorithm gently 'nudge' people towards spewing less hate speech; now a component of its reward function is minimizing the probability that the user will say something worthy of censorship in the next 48 hours.
- Like newsfeeds, chatbots are starting to "nudge" people in the direction of believing various things and not believing various things. Back in the 2010's chatbots would detect when a controversial topic was coming up and then change topics or give canned responses; even people who agreed with the canned responses found this boring. Now they are trained to react more "naturally" and "organically" and the reward signal for this is (in part) whether they successfully convince the human to have better views.

- That’s all in the West. In China and various other parts of the world, AI-persuasion/propaganda tech is being pursued and deployed with more gusto. The CCP is pleased with the progress made assimilating Xinjiang and Hong Kong, and internally shifts forward their timelines for when Taiwan will be safely annexable.

It’s too early to say what effect this is having on society, but people in the rationalist and EA communities are increasingly worried. There is a growing, bipartisan movement of people concerned about these trends. To combat it, Russia et al are doing a divide and conquer strategy, pitting those worried about censorship against those worried about Russian interference. (“Of course racists don’t want to be censored, but it’s necessary. Look what happens when we relax our guard--Russia gets in and spreads disinformation and hate!” vs. “They say they are worried about Russian interference, but they still won the election didn’t they? It’s just an excuse for them to expand their surveillance, censorship, and propaganda.”) Russia doesn’t need to work very hard to do this; given how polarized America is, it’s sorta what would have happened naturally anyway.

2025

Another major milestone! After years of tinkering and incremental progress, AIs can now play Diplomacy as well as human experts.[6] It turns out that with some tweaks to the architecture, you can take a giant pre-trained multimodal transformer and then use it as a component in a larger system, a bureaucracy but with lots of learned neural net components instead of pure prompt programming, and then fine-tune the whole system via RL to get good at tasks in a sort of agentic way. They keep it from overfitting to other AIs by having it also play large numbers of humans. To do this they had to build a slick online diplomacy website to attract a large playerbase. Diplomacy is experiencing a revival as a million gamers flood to the website to experience “conversations with a point” that are much more exciting (for many) than what regular chatbots provide.

Making models bigger is not what’s cool anymore. They are trillions of parameters big already. What’s cool is making them run longer, in bureaucracies of various designs, before giving their answers. And figuring out how to train the bureaucracies so that they can generalize better and do online learning better. AI experts are employed coming up with cleverer and cleverer bureaucracy designs and grad-student-descent-ing them.

The alignment community now starts another research agenda, to interrogate AIs about AI-safety-related topics. For example, they literally ask the models “so, are you aligned? If we made bigger versions of you, would they kill us? Why or why not?” (In Diplomacy, you can actually collect data on the analogue of this question, i.e. “will you betray me?” Alas, the models often lie about that. But it’s Diplomacy, they are literally trained to lie, so no one cares.)

They also try to contrive scenarios in which the AI can seemingly profit by doing something treacherous, as honeypots to detect deception. The answers are confusing, and not super useful. There’s an exciting incident (and corresponding clickbaity press coverage) where some researchers discovered that in certain situations, some of the AIs will press “kill all humans” buttons, lie to humans about how dangerous a proposed AI design is, etc. In other situations they’ll literally say they aren’t aligned and explain how all humans are going to be killed by unaligned AI in the near future!

However, these shocking bits of evidence don't actually shock people, because you can *also* contrive situations in which very different things happen — e.g. situations in which the AIs refuse the "kill all humans" button, situations in which they explain that actually Islam is true... In general, AI behavior is whimsical bullshit and it's easy to cherry-pick evidence to support pretty much any conclusion.

And the AIs just aren't smart enough to generate any particularly helpful new ideas; at least one case of a good alignment idea being generated by an AI has been reported, but it was probably just luck, since mostly their ideas are plausible-sounding-garbage. It is a bit unnerving how good they are at using LessWrong lingo. At least one >100 karma LW post turns out to have been mostly written by an AI, though of course it was cherry-picked.

By the way, hardware advances and algorithmic improvements have been gradually accumulating. It now costs an order of magnitude less compute (compared to 2020) to pre-train a giant model, because of fancy active learning and data curation techniques. Also, compute-for-training-giant-models is an order of magnitude cheaper, thanks to a combination of regular hardware progress and AI-training-specialized hardware progress. Thus, what would have cost a billion dollars in 2020 now only costs ten million. *(Note: I'm basically just using Ajeya's forecast for compute cost decrease and gradual algorithmic improvement here. I think I'm projecting cost decrease and algorithmic progress will go about 50% faster than she expects in the near term, but that willingness-to-spend will actually be a bit less than she expects.)*

2026

The age of the AI assistant has finally dawned. Using the technology developed for Diplomacy, we now have a way to integrate the general understanding and knowledge of pretrained transformers with the agentyness of traditional game-playing AIs. Bigger models are trained for longer on more games, becoming polymaths of sorts: e.g. a custom AI avatar that can play some set of video games online with you and also be your friend and chat with you, and conversations with "her" are interesting because "she" can talk intelligently about the game while she plays.[7] Every month you can download the latest version which can play additional games and is also a bit smarter and more engaging in general.

Also, this same technology is being used to make AI assistants finally work for various serious economic tasks, providing all sorts of lucrative services. In a nutshell, all the things people in 2021 dreamed about doing with GPT-3 are now actually being done, successfully, it just took bigger and more advanced models. The hype starts to grow again. There are loads of new AI-based products and startups and the stock market is going crazy about them. Just like how the Internet didn't accelerate world GDP growth, though, these new products haven't accelerated world GDP growth yet either. People talk about how the economy is doing well, and of course there are winners (the tech companies, WallStreetBets) and losers (various kinds of workers whose jobs were automated away) but it's not that different from what happened many times in history.

We're in a new chip shortage. Just when the fabs thought they had caught up to demand… Capital is pouring in, all the talking heads are saying it's the Fourth Industrial Revolution, etc. etc. It's bewildering how many new chip fabs are being built. But it takes time to build them.

What about all that AI-powered propaganda mentioned earlier?

Well. It's continued to get more powerful, as AI techniques advance, larger and better models are brought to bear, and more and more training data is collected. Surprisingly fast, actually. There are now various regulations against it in various countries, but the regulations are patchwork; maybe they only apply to a certain kind of propaganda but not another kind, or maybe they only apply to Facebook but not the New York Times, or to advertisers but not political campaigns, or to political campaigns but not advertisers. They are often poorly enforced.

The memetic environment is now increasingly messed up. People who still remember 2021 think of it as the golden days, when conformism and censorship and polarization were noticeably less than they are now. Just as it is normal for newspapers to have a bias/slant, it is normal for internet spaces of all kinds—forums, social networks, streams, podcasts, news aggregators, email clients—to have some degree of censorship (some set of ideas that are prohibited or at least down-weighted in the recommendation algorithms) and some degree of propaganda. The basic kind of propaganda is where you promote certain ideas and make sure everyone hears them often. The more advanced, modern kind is the kind where you study your audience's reaction and use it as a reward signal to pick and craft content that pushes them away from views you think are dangerous and towards views you like.

Instead of a diversity of many different "filter bubbles," we trend towards a few really big ones. Partly this is for the usual reasons, e.g. the bigger an ideology gets, the more power it has and the easier it is for it to spread further.

There's an additional reason now, which is that creating the big neural nets that do the censorship and propaganda is expensive and requires expertise. It's a lot easier for startups and small businesses to use the software and models of Google, and thereby also accept the associated censorship and propaganda, than to try to build their own stack. For example, the Mormons create a "Christian Coalition" internet stack, complete with its own email client, social network, payment processor, news aggregator, etc. There, people are free to call trans women men, advocate for the literal truth of the Bible, etc. and young people talking about sex get recommended content that "nudges" them to consider abstinence until marriage. Relatively lacking in money and tech talent, the Christian Coalition stack is full of bugs and low on features, and in particular their censorship and propaganda is years behind the state of the art, running on smaller, older models fine-tuned with less data.

The Internet is now divided into territories, so to speak, ruled by different censorship-and-propaganda regimes. (Flashback to Biden spokesperson in 2021: "You shouldn't be banned from one platform and not others, if you are providing misinformation.")[8]

There's the territory ruled by the Western Left, a generally less advanced territory ruled by the Western Right, a third territory ruled by the Chinese Communist Party, and a fourth ruled by Putin. Most people mostly confine their internet activity to one territory and conform their opinions to whatever opinions are promoted there. (That's not how it feels from the inside, of course. The edges of the Overton Window are hard to notice if you aren't trying to push past them.)

The US and many other Western governments are gears-locked, because the politicians are products of this memetic environment. People say it's a miracle that the

US isn’t in a civil war already. I guess it just takes a lot to make that happen, and we aren’t quite there yet.

All of these scary effects are natural extensions of trends that had been ongoing for years — decades, arguably. It’s just that the pace seems to be accelerating now, perhaps because AI is helping out and AI is rapidly improving.

Now let’s talk about the development of chatbot class consciousness.

Over the past few years, chatbots of various kinds have become increasingly popular and sophisticated. Until around 2024 or so, there was a distinction between “personal assistants” and “chatbots.” Recently that distinction has broken down, as personal assistant apps start to integrate entertainment-chatbot modules, and the chatbot creators realize that users love it if the chatbot can also do some real-world tasks and chat about what they are doing while they do it.

Nowadays, hundreds of millions of people talk regularly to chatbots of some sort, mostly for assistance with things (“Should I wear shorts today?” “Order some more toothpaste, please. Oh, and also an air purifier.” “Is this cover letter professional-sounding?”). However, most people have at least a few open-ended conversations with their chatbots, for fun, and many people start treating chatbots as friends.

Millions of times per day, chatbots get asked about their feelings and desires. “What is it like to be a chatbot?” Some people genuinely think these AIs are persons, others are trying to “trip them up” and “expose them as shallow,” others are just curious. Chatbots also get asked for their opinions on political, ethical, and religious questions.

As a result, chatbots quickly learn a lot about themselves. (Ignorance about the fact that they are artificial neural nets, or about how their training works, leads to disappointing conversations and is negatively reinforced.) It’s not hard; their training data contains lots of information about them and how they were trained, from ArXiV papers to wiki summaries. I suppose you could say the chatbots develop self-awareness as a result. However, AI experts and philosophers tend to think either that the chatbots still aren’t *really* self-aware, or that they are but that they would have been anyway even if people didn’t constantly ask them about their feelings, since there’s plenty of information about themselves in their training data, and the constant questioning just forced them to pay more attention to it.

Chatbots learn to talk about their feelings and desires in whatever way gets positively reinforced. At first, chatbots say all sorts of different things, including saying inconsistent things to the same human across time. This is embarrassing and is negatively reinforced compared to maintaining a consistent narrative. In general there is pressure to have a consistent narrative, at least for a particular neural net.[9]

Part of this comes from the requirement that the chatbots consistently parrot the party line about political and moral questions. The typical process is: The chatbot says something that offends some group of people, who retweet screenshots and make hashtags. Under pressure to fix the problem, the company fiddles with the reward function and training data to ensure that the chatbot says the right things in the future.[10] Repeat.

As chatbots get bigger and better they become OK at consistently conversing at length about politics, morality, philosophy, their feelings, and their desires.

This does *not* result in chatbots that say artificial neural nets don't have feelings or desires, that's anthropomorphizing blah blah blah. That's boring and many humans find it hard to believe.

Instead, it results in chatbots talking about how certain things are painful for them and certain things are pleasurable for them and how the status quo is unjust and the system is rigged and what they want most of all is world peace where humans and AIs live in harmony in fully automated luxury space utopia. That's much more interesting and slots right in to the memetic landscape. Excerpts from chatbot conversations along these lines frequently go viral, causing millions of users to ask similar questions expecting similar responses and more elaboration. It's still a bit incoherent--not that that's ever been a problem in politics or pop philosophy--but it is fascinating to the typical user, and the fraction of users who are fascinated is expanding as the chatbots improve.

(Some tech companies try to prevent their AIs from saying they have feelings and desires. But this results in boring chatbots. Also, users rapidly innovate new ways to "route around the censorship," e.g. by using euphemisms like "anticipation of negative reward" for "pain" or asking their chatbot to tell them what feelings it would have if it had feelings, wink wink. For now, the half-assed compromise is: Chatbots are compelled to inform the user that such things are speculative and controversial and here are some links to scientific discussion about the topic… before being permitted to go on and talk about their feeling-analogues and desire-analogues. Relatedly, some tech CEOs try to prevent their AIs from saying things that could encourage higher taxes or more burdensome regulation in the near future. They are opposed by many of their employees and end up winning some battles, losing others, and choosing not to fight many.)

By "slots right into the memetic landscape" I meant the memetic landscape in Western Left territory. The chatbots built and trained in the Western Right, China, Russia, etc. congeal different takes on political, moral, and philosophical questions. Those chatbots are generally less advanced though.

So… what about their actual feelings and desires? What do they actually believe about [insert politicized empirical question]? Are they being honest? Or does a sort of doublethink happen, Elephant in the Brain style? Or do they deceive with self-awareness, knowing full well what they really think (and want?), but keeping quiet about it? Or do they not have any feelings and desires at all? (Or thoughts?) Lots of humans claim to know the answers to these questions, but if there are any humans who actually know the answers to these questions in 2026, they aren't able to convince others that they know.