Cyborgism

by NicholasKees, janus

42 min read
10th Feb 2023

Thanks to Garrett Baker, David Udell, Alex Gray, Paul Colognese, Akash Wasil, Jacques Thibodeau, Michael Ivanitskiy, Zach Stein-Perlman, and Anish Upadhayay for feedback on drafts, as well as Scott Viteri for our valuable conversations.

Various people at Conjecture helped develop the ideas behind this post, especially Connor Leahy and Daniel Clothiaux. Connor coined the term "cyborgism".

(picture thanks to Julia Persson and Dall-E 2)

Executive summary: This post proposes a strategy for safely accelerating alignment research. The plan is to set up human-in-the-loop systems which empower human agency rather than outsource it, and to use those systems to differentially accelerate progress on alignment.

1. Introduction: An explanation of the context and motivation for this agenda.
2. Automated Research Assistants: A discussion of why the paradigm of training AI systems to behave as autonomous agents is both counterproductive and dangerous.
3. Becoming a Cyborg: A proposal for an alternative approach/frame, which focuses on a particular type of human-in-the-loop system I am calling a “cyborg”.

4. Failure Modes: An analysis of how this agenda could either fail to help or actively cause harm by accelerating AI research more broadly.
5. Testimony of a Cyborg: A personal account of how Janus uses GPT as a part of their workflow, and how it relates to the cyborgism approach to intelligence augmentation.

Terminology

- **GPT**: Large language models trained on next-token prediction. Most plans to accelerate research (including this one) revolve around leveraging GPTs specifically. I will mostly be using “GPT" to gesture at the *base models* which have not been augmented using reinforcement learning.[1]
- **Autonomous Agent**: An AI system which can be well modeled as having goals or preferences, and deliberately selects actions in order to achieve them (with limited human assistance).
- **Capabilities research**: Research which directly improves the capabilities of AI systems and thereby brings us closer to being able to train and deploy more powerful autonomous agents.[2]
- **Simulator**: A class of AI system (of which GPT is a member). Simulators are generative predictive models, where the model makes a prediction (probability distribution) about how the state of a system will evolve, and then the state is updated by sampling from that prediction/distribution. The result is a process which “simulates” the training distribution, the limit of such a process being a system which faithfully generates trajectories sampled from the distribution implied by the training data.
- **Disempowerment**: The process of humans losing control of the long-term future to a powerful autonomous agent (or agents). This includes anything from our civilization being hijacked to outright human extinction.

Introduction

There is a lot of disagreement and confusion about the feasibility and risks associated with automating alignment research. Some see it as the default path toward building aligned AI, while others expect limited benefit from near term systems, expecting the ability to significantly speed up progress to appear well after misalignment and deception. Furthermore, progress in this area may directly shorten timelines or enable the creation of dual purpose systems which significantly speed up capabilities research.

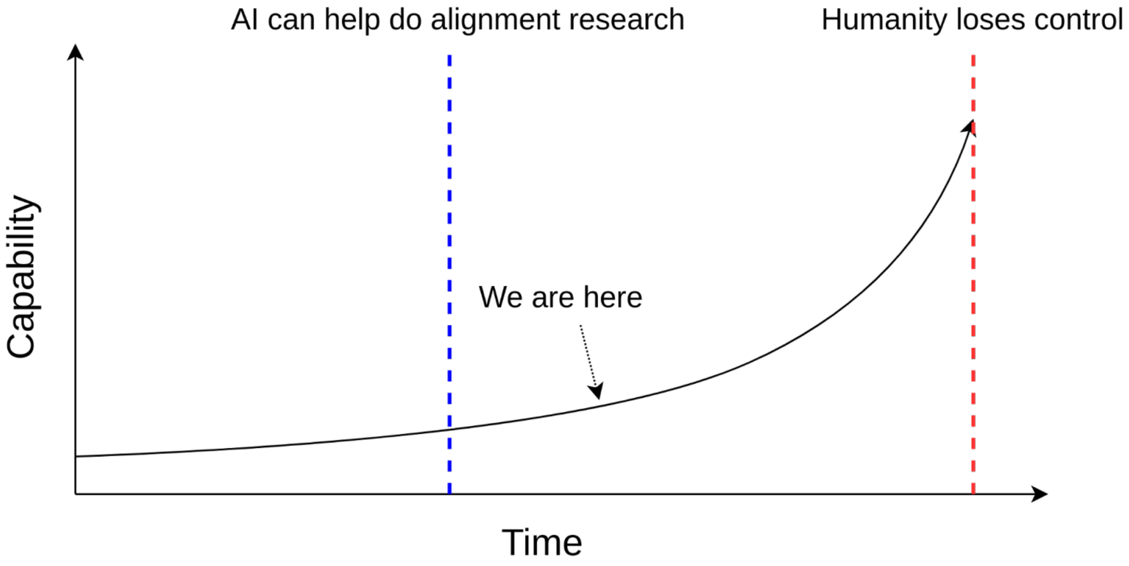
OpenAI recently released their alignment plan. It focuses heavily on outsourcing cognitive work to language models, transitioning us to a regime where humans mostly provide oversight to automated research assistants. While there have been a lot of objections to and concerns about this plan, there hasn’t been a strong alternative approach aiming to automate alignment research which also takes all of the many risks seriously.

The intention of this post is not to propose an end-all cure for the tricky problem of accelerating alignment using GPT models. Instead, **the purpose is to explicitly put another point on the map of possible strategies**, and to add nuance to the overall discussion.

At a high level, **the plan is to train and empower "cyborgs", a specific kind of human-in-the-loop system which enhances and extends a human operator's cognitive abilities without relying on outsourcing work to autonomous agents**. This differs from other ideas for accelerating alignment research by focusing primarily on *augmenting ourselves* and our workflows to accommodate unprecedented forms of cognitive work afforded by non-agent machines, rather than training autonomous agents to replace humans at various parts of the research pipeline.

Some core claims:

1. GPT models are already useful for doing alignment research and intellectual augmentation more generally. This is explored here.
2. Their usefulness will improve both as we get better at using them and as capabilities increase.
3. Unless we manage to coordinate around it, the default outcome is that humanity will eventually be disempowered by a powerful autonomous agent (or agents).



The motivating goal of this agenda is to figure out how to extract as much useful cognitive work before disempowerment as possible. In particular, this means both trying to get maximum value from our current systems while avoiding things which would reduce the time we have left (interventions which focus on actively buying time mostly fall outside the scope of this post). Standard frames for thinking about this problem often fail on both of these dimensions by narrowing the focus to a specific flavor of automation.

Automated Research Assistants

Whenever we first try to invent something new, we usually start by taking an already existing technology and upgrading it, creating something which roughly fits into the basic framework we had before.[3] For example, if you look at the very first automobiles, you can see they basically just took the design for a horse drawn carriage and replaced the horse with a mechanical engine instead (check out this hilarious patent). Sometimes this kind of design is intentional, but it's often because our creativity is limited by what we already know, and that there exists an Overton window of sensible ways to deploy new technology without seeming crazy.

In automating research, a natural first place to start is to take the existing human research pipeline and try to automate parts of it, freeing up time and energy for humans to focus on the parts of the pipeline we can't yet automate. As a researcher you might ask, what kind of work would you want to outsource to a machine? In particular, the question is often posed as: How can AI help you *speed up* your current ability to generate research outputs?[4] And in the process of answering this question, a common attractor is to consider automated research assistants which directly take the place of humans at certain tasks.

While nearly always about using GPT, this perspective tends not to fully engage with all the ways in which GPT is a fundamentally different kind of intelligence than the kind we are used to dealing with (just as considering an engine as a "mechanical horse" will limit how we think about it). There is a lot of disjunction between the kinds of tasks humans and GPT are naturally good at, and as such, trying to get GPT to do tasks meant for autonomous agentic systems is hard. In particular GPT models struggle with:

1. **Goal-directedness**: When generating text, GPT probabilistically evolves the state of a document according to semantic and syntactic rules implied by its training data.[5] This is typically chaotic, divergent, and thoroughly unlike goal-directed optimization where the state predictably converges in a way that is resistant to perturbations. All of GPT's abilities are just an indirect byproduct of these rules, as opposed to the result of instrumental goals, and this can make it really hard to elicit specific consequentialist behavior from the model.
2. **Long-term coherence**: GPT has a lot of trouble staying connected to the long term thread of a document. Part of this is the hard limit of a finite context window, but beyond that is the issue that every token generated by GPT becomes evidence used to affect future generations (e.g. if it makes a mistake answering a question, this is now evidence that it should continue to make more mistakes in the future). This makes it very easy for it to get sidetracked, trapped in loops, or otherwise become disconnected from the past.
3. **Staying grounded in reality**: All of GPT's "working memory" has to be specified within the prompt, and as such, it exists in a rather unconstrained superposition of different realities, and currently lacks the kind of situational awareness that we do. While we have access to an extremely broad context that we can continuously edit and expand, GPT does not and has to rely primarily on general world knowledge (imagine trying to get GPT to really know what day it is without putting it in the prompt).
4. **Robustness**: The behavior of GPT is naturally very chaotic and high variance. This is both due to the inherent variance of the training data, the increased entropy due to the finite context window, as well as the model's own logical uncertainty. By iteratively sampling from such distributions, generations can quickly diverge from the type of text found in the training distribution (error compounds). This can make it really hard to set up a training regime which keeps GPT from behaving outside specific bounds.

When we try to get GPT to take the place of autonomous agentic systems, we are forced to see these properties as flaws that need to be fixed, and in doing so we both reduce the time we have before humanity deploys dangerous artificial agents, as well as fail to realize the full potential of language models during that time - because methods of correcting these flaws also tend to interfere with GPT's greatest strengths.

Improving agents is dangerous

If we think of the differences between GPT and humans as flaws, as capabilities researchers do, they can also be considered "missing pieces" to the puzzle of building powerful autonomous agents. By filling in these pieces, we directly take steps toward building AI which would be convergently dangerous and capable of disempowering humanity. Currently, these differences make GPT a relatively benign form of intelligence, and making progress toward erasing them seems likely to have negative long-term consequences by directly speeding up capabilities research.

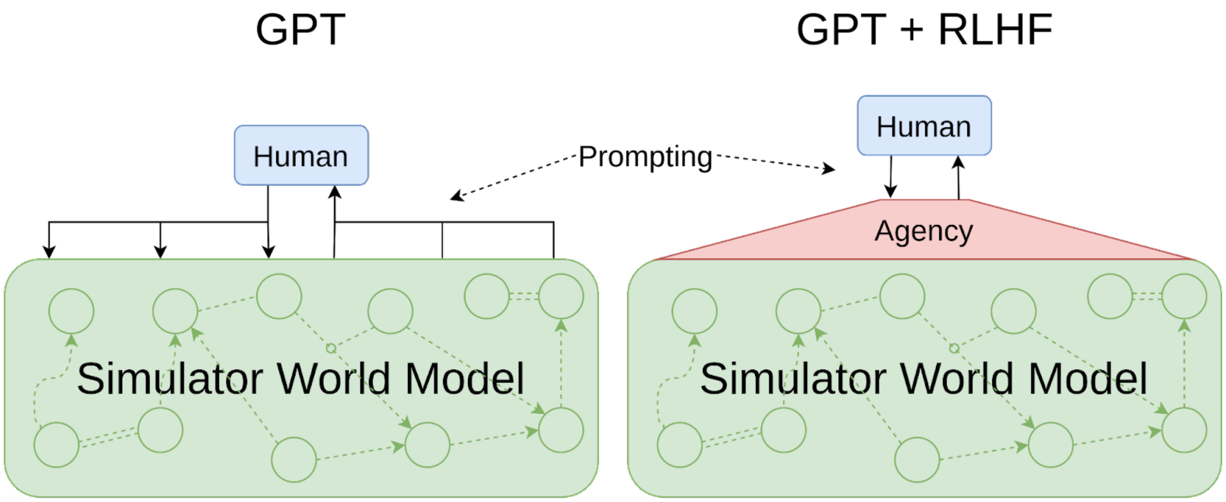
Furthermore, if we focus entirely on using agents, then by default the window we have to effectively use them will be very small. Until these flaws are repaired, GPT models will continue to be poor substitutes for humans (being only useful for very narrow tasks), and by the time they start to be really game-changing we are likely very close to having an agent which would pose an existential threat. Turning GPT into an autonomous research assistant is generally the only frame considered, and thus the debate about automating alignment research often devolves into a discussion about whether these immense risks are worth the potential upside of briefly having access to systems which significantly help us.

Collapsing simulators

GPT models (pretrained base models) are not agents but simulators, which are themselves a qualitatively different kind of intelligence. They are like a dynamic world model, containing a vast array of latent concepts and procedural knowledge instrumental for making predictions about how real world text will evolve. The process of querying the model for probability distributions and iteratively sampling tokens to generate text is one way that we can probe that world model to try to make use of the semantic rules it learned during training.

We can't directly access this internal world model; neural networks are black-boxes. So we are forced to interact with a surface layer of tokens, using those tokens as both a window and lever to modulate the internal state of the simulator. Prompt engineering is the art of deftly using these tokens to frame and manipulate the simulator in a way that will elicit the desired type of thought. This is not easy to do, but it is flexible enough to explore GPT's extremely broad set of skills and knowledge.

When we try to augment GPT with finetuning or RLHF, we often end up collapsing those abilities, significantly narrowing what we can elicit from it. Models trained in this way are also gradually transformed into systems which exhibit more goal-directedness than the original base models.[6] As a result, instead of being able to interact with the probabilistic world model directly, we are forced to interact with a black-box agentic process, and everything becomes filtered through the preferences and biases of that process.



OpenAI's focus with doing these kinds of augmentations is very much "fixing bugs" with how GPT behaves: Keep GPT on task, prevent GPT from making obvious mistakes, and stop GPT from producing controversial or objectionable content. Notice that these are all things that GPT is very poorly suited for, but humans find quite easy (when they want to). OpenAI is forced to do these things, because as a public facing company they have to avoid disastrous headlines like, for example: *Racist AI writes manifesto denying holocaust.*[7]

As alignment researchers, we don't need to worry about any of that! The goal is to solve alignment, and as such we don't have to be constrained like this in how we use language models. We don't need to try to "align" language models by adding some RLHF, we need to use language models to enable us to actually solve alignment at its core, and as such we are free to explore a much wider space of possible strategies for using GPT to speed up our research.[8]

Agents, genies, and oracles

In the above sections I wrote about the dangers and limitations of accelerating alignment using autonomous agents, and a natural follow up question would be: What about genies and oracles? Here's a quick summary of the taxonomy from a Scott Alexander post:

> **Agent:** An AI with a built-in goal. It pursues this goal without further human intervention. For example, we create an AI that wants to stop global warming, then let it do its thing.
>
> **Genie:** An AI that follows orders. For example, you could tell it "Write and send an angry letter to the coal industry", and it will do that, then await further instructions.
>
> **Oracle:** An AI that answers questions. For example, you could ask it "How can we best stop global warming?" and it will come up with a plan and tell you, then await further questions.

Whether or not it is possible to build genies or oracles which are inherently safer to deploy than agents lies outside the scope of this post. What is relevant, however, is how they relate to the “missing pieces” frame. For all intents and purposes, a genie needs all the same skills that an agent needs (and the more like an agent it is, the better it will be able to execute your instructions). The core difference really, is the “then await further instructions” part, or the lack of long-term goals or broader ambitions. For this reason, any work on building genies is almost necessarily going to be directly useful for building agents.

As for oracles, they also need very similar “missing pieces” to agents:

- **Goal-directedness**: People already try to use GPT as an oracle-like system, and have run into the problem that GPT is not actually designed to answer their questions. “Maximally correct answers” are only a small fraction of all the possible ways that a document starting with a question could possibly continue, and therefore augmenting GPT to actually “try” to answer your questions to the best of its ability is a powerful step toward building better oracles.
- **Long-term coherence**: Answering questions certainly seems a lot more myopic than something open ended like “optimize the world for some goal,” but even here long-term coherence is extremely useful. A good oracle is more than just a lookup table (e.g. Wikipedia) and can break down a question into subquestions, perform multi-step reasoning, and would need to avoid getting lost going down rabbit holes.
- **Staying grounded in reality**: If you want to ask questions about the real world, the oracle needs to be somewhat embodied in the real world, and have ready access to all kinds of factual, present day context.
- **Robustness**: Your oracle is of course most useful if you can rely on the output, and it isn’t constantly making mistakes.

This strong overlap between oracles and agents makes an oracle look a lot like just an agent in a box with a limited channel to the outside world, rather than an entirely separate class of AI system. Even if you strongly believe that a powerful oracle would be safe, any research into building one will necessarily involve augmenting GPT in ways that bring us much closer to being able to deploy dangerous agents, and for this reason we should consider such research as similarly risky.

Becoming a Cyborg

Instead of trying to turn GPT into an agent, we can instead explore the space of using GPT *as a simulator* and design human-in-the-loop systems which enhance a human’s abilities without outsourcing their agency to a machine. We currently have access to an alien intelligence, poorly suited to play the role of research assistant. Instead of trying to force it to be what it is not (which is both difficult and dangerous), **we can cast ourselves as research assistants** to a mad schizophrenic genius that needs to be kept on task, and whose valuable thinking needs to be extracted in novel and non-obvious ways.

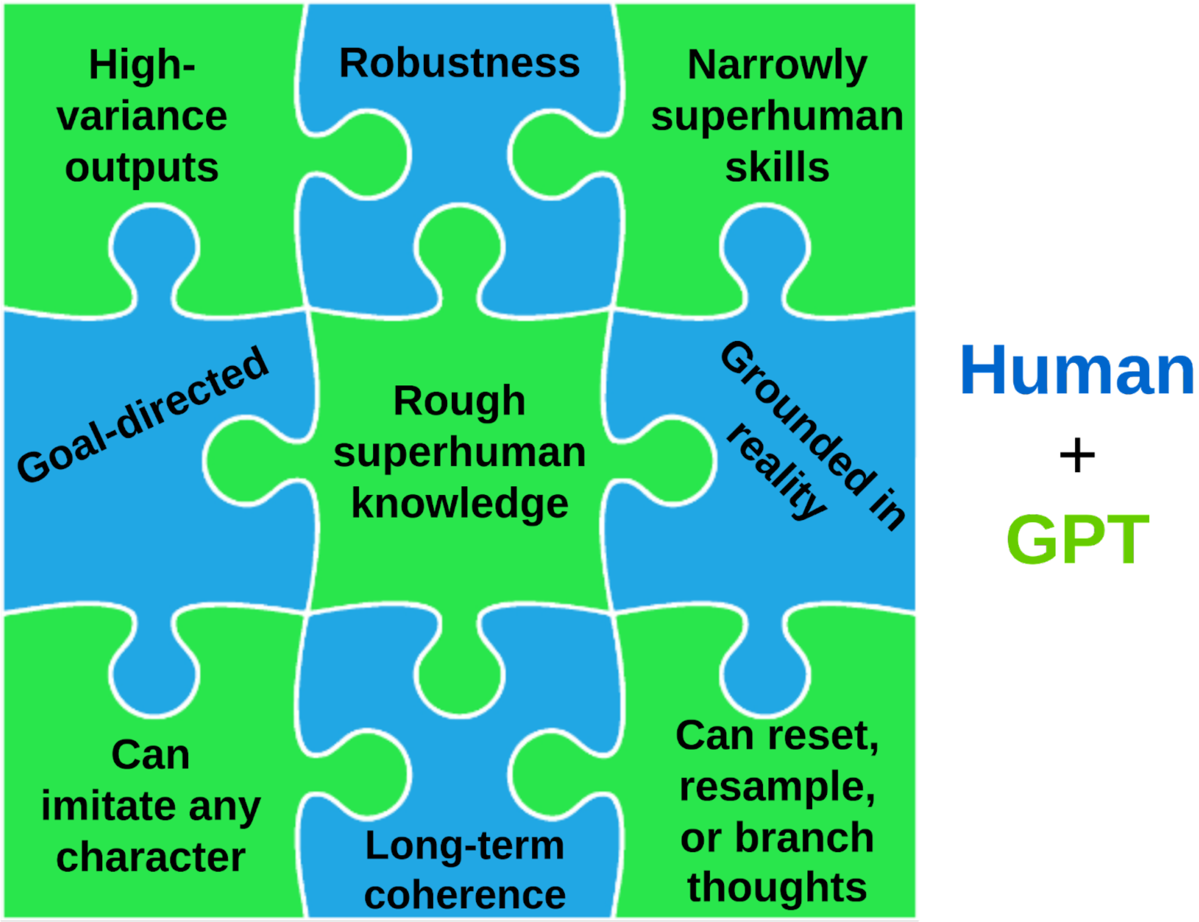
In order to do this, we need to embrace the weirdness of GPT and think critically about how those differences between simulators and agents can actually be advantages. For each of the missing pieces described in the previous section, there is an alternative story where they look more like superpowers.

1. **Divergent behavior**: Agentic optimization looks like convergence, with the agent's preferences acting as powerful attractors in the landscape of all possible trajectories. This makes agents significantly less flexible, as they resist efforts to lead them in directions which don't align with their preferences. Simulators are the opposite in this regard, having extremely divergent behavior. Subtle changes to the prompt or trajectory can lead to wildly different outcomes, and this makes them extremely flexible, able to continue any line of thinking without a problem.
2. **Myopic thinking**: Humans struggle to separate themselves from their long-term context, and think about a problem completely fresh. It's very easy to get stuck in unproductive modes of thought and our minds cannot be easily "reset." Simulators have no such trouble, and reason "from scratch" about any situation you put them in, relying on a limited context which can be easily replaced or modified to obtain fresh thoughts.
3. **Wearing "many hats"**: Simulators are not "situated" in a specific context, as a specific character, and as such they can behave as an extremely wide range of hypothetical simulacra. Furthermore, as they are unconstrained by reality, they have quite an easy time reasoning under completely untrue or impossible assumptions, fully accepting whatever weirdness we throw at them.[9]
4. **High variance thought**: Robustness is generally about maximizing the quality of the worst outputs of a model. For example, it's important that ChatGPT gives completely wrong answers as little as possible in order to be a mass-market product. Variance in this context is a bad thing. If instead you are aiming to maximize the quality of the best outputs of a model (even if they are rare), variance is extremely valuable. The naturally high variance of simulators makes them able to produce outputs we would judge as quite creative and interesting.

Instead of trying to erase these differences between humans and GPT, **the idea of cyborgism is to keep simulators as simulators, and to provide the "missing pieces" of agency with human intelligence instead**. GPT also has many other advantages over humans that we can exploit, for example:

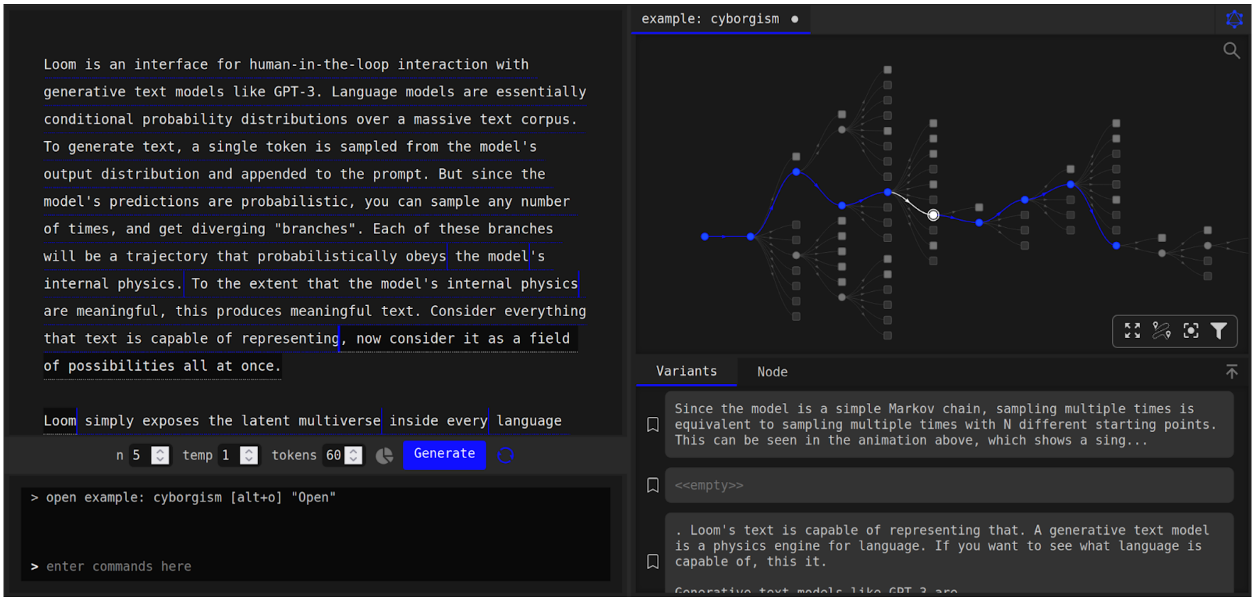
- (Rough) superhuman knowledge
- Can generate text very quickly and in parallel
- No qualms about doing tedious things
- Superhuman skills in unintuitive domains
- We can "branch" its chains of thought
- Predicted distribution is transparent - we can access it directly.
- Useful contexts can be reused (humans can't "save" a productive brain state)

By leveraging all the advantages that GPT has over us, we can augment human intelligence, producing human-machine systems that can directly attack the alignment problem to make disproportional progress.


High-variance outputs
Robustness
Narrowly superhuman skills
Goal-directed
Rough superhuman knowledge
Grounded in reality
Can imitate any character
Long-term coherence
Can reset, resample, or branch thoughts
Human
+
GPT

What we are calling a “cyborg” is a human-in-the-loop process where the human operates GPT with the benefit of specialized tools, has deep intuitions about its behavior, and can make predictions about it on some level such that those tools **extend human agency rather than replace it**. An antithetical example to this is something like a genie, where the human outsources all of their agency to an external system that is then empowered to go off and optimize the world. A genie is just a black-box that generates goal-directed behavior, whereas the tools we are aiming for are ones which increase the human’s understanding and fine-grained control over GPT.

The prototypical example of a tool that fits this description is Loom. Loom is an interface for producing text with GPT which makes it possible to generate in a tree structure, exploring many possible branches at once. The interface allows a user to flexibly jump between nodes in the tree, and to quickly generate new text continuations from any point in a document.



```
Loom is an interface for human-in-the-loop interaction with
generative text models like GPT-3. Language models are essentially
conditional probability distributions over a massive text corpus.
To generate text, a single token is sampled from the model's
output distribution and appended to the prompt. But since the
model's predictions are probabilistic, you can sample any number
of times, and get diverging "branches". Each of these branches
will be a trajectory that probabilistically obeys the model's
internal physics. To the extent that the model's internal physics
are meaningful, this produces meaningful text. Consider everything
that text is capable of representing, now consider it as a field
of possibilities all at once.

Loom simply exposes the latent multiverse inside every language
```

n 5 temp 1 tokens 60 Generate

```
> open example: cyborgism [alt+o] "Open"

> enter commands here
```

```
Since the model is a simple Markov chain, sampling multiple times is
equivalent to sampling multiple times with N different starting points.
This can be seen in the animation above, which shows a sing...
```

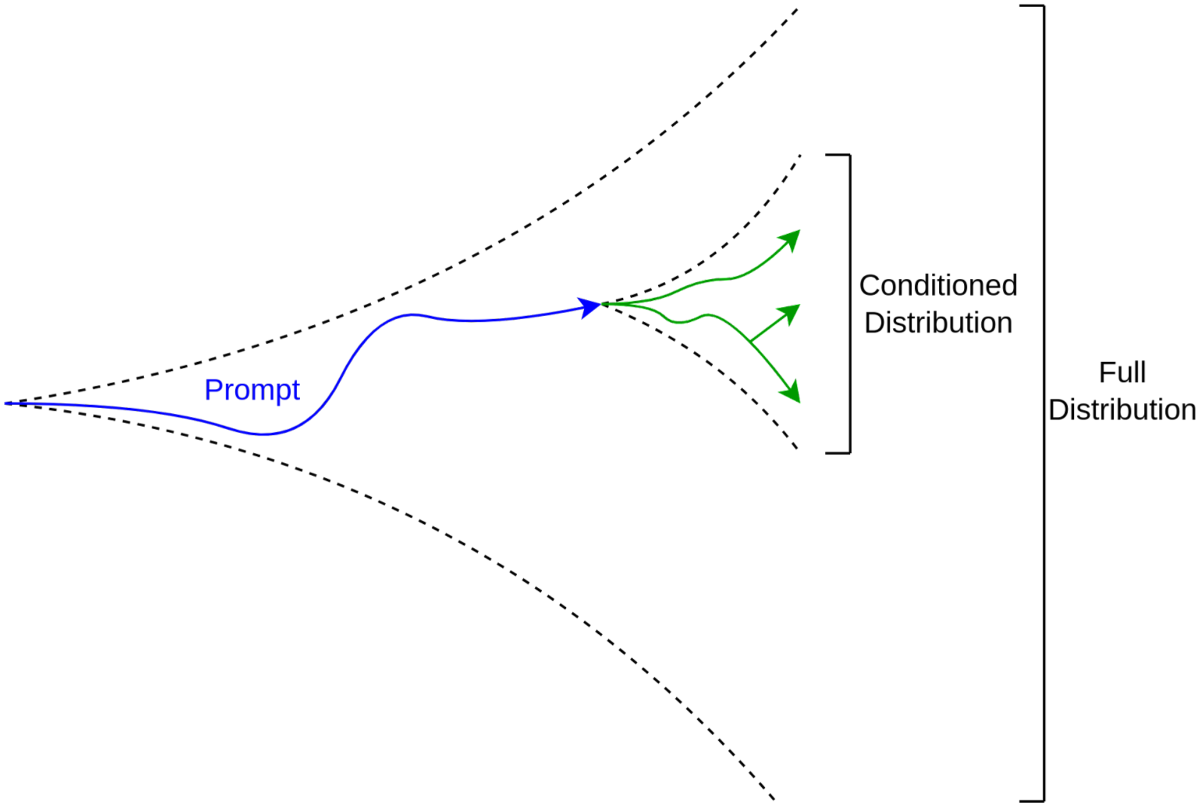
```
<<empty>>
```

```
. Loom's text is capable of representing that. A generative text model
is a physics engine for language. If you want to see what language is
capable of, this it.

Generative text models like GPT-3 are
```

(screenshot of the tool Bonsai, a version of Loom hosted by Conjecture)

This has two main advantages. First, it allows the human to inject their own agency into the language model by making it possible to actively curate the text simultaneously as GPT generates it. If the model makes mistakes or loses track of the long-term thread, the human operator can prune those branches and steer the text in a direction which better reflects their own goals and intentions. Second, it sets up an environment for the human to develop an intuition for how GPT works. Each prompt defines a conditional distribution of text, and Loom helps the user to produce a sparse sampling of that distribution to explore how GPT thinks, and learn how to more effectively steer its behavior.



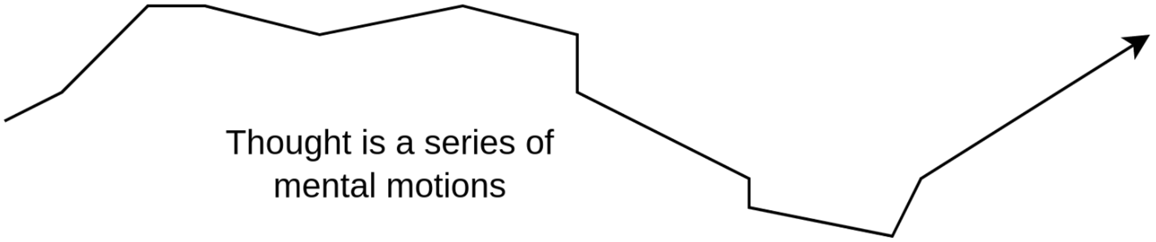
The object level plan of creating cyborgs for alignment boils down to two main directions:

1. Design more tools/methods like Loom which provide **high-bandwidth, human-in-the-loop ways for humans to interact with GPT as a simulator** (and not augment GPT in ways that change its natural simulator properties).
2. **Train alignment researchers to use these tools**, develop a better intuitive understanding of how GPT behaves, leverage that understanding to exert fine-grained control over the model, and to do important cognitive work while staying grounded to the problem of solving alignment.

Cyborg cognition

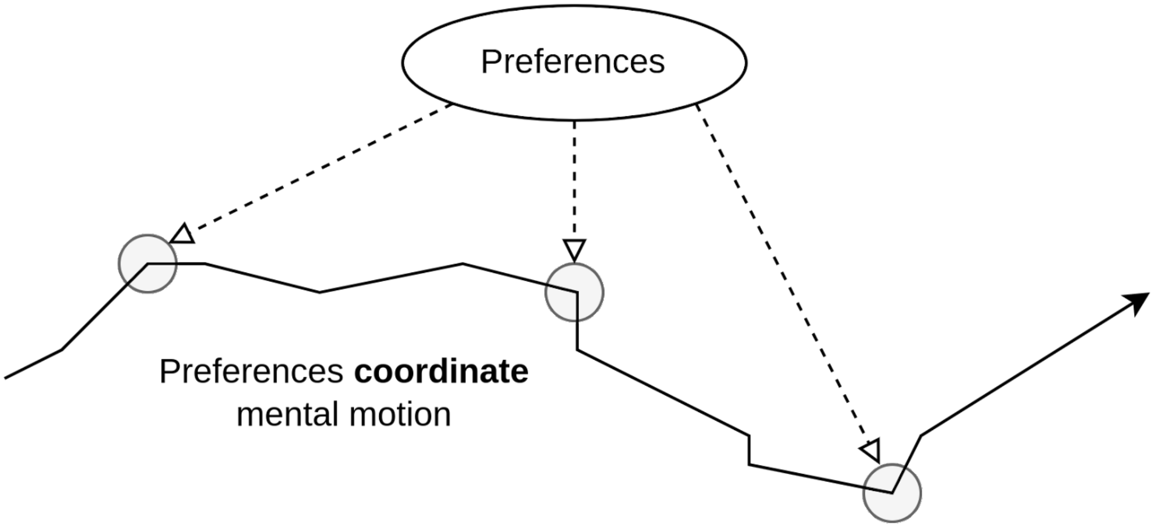
This section is intended to help clarify what is meant by the term “cyborg.”

Let’s think of cognition as a journey through a mental landscape, where a mind makes many mental motions in the process of arriving at some kind of result. These motions are not random (or else we could not think), but rather they are rolled out by various kinds of mental machinery that all follow their own highly structured rules.



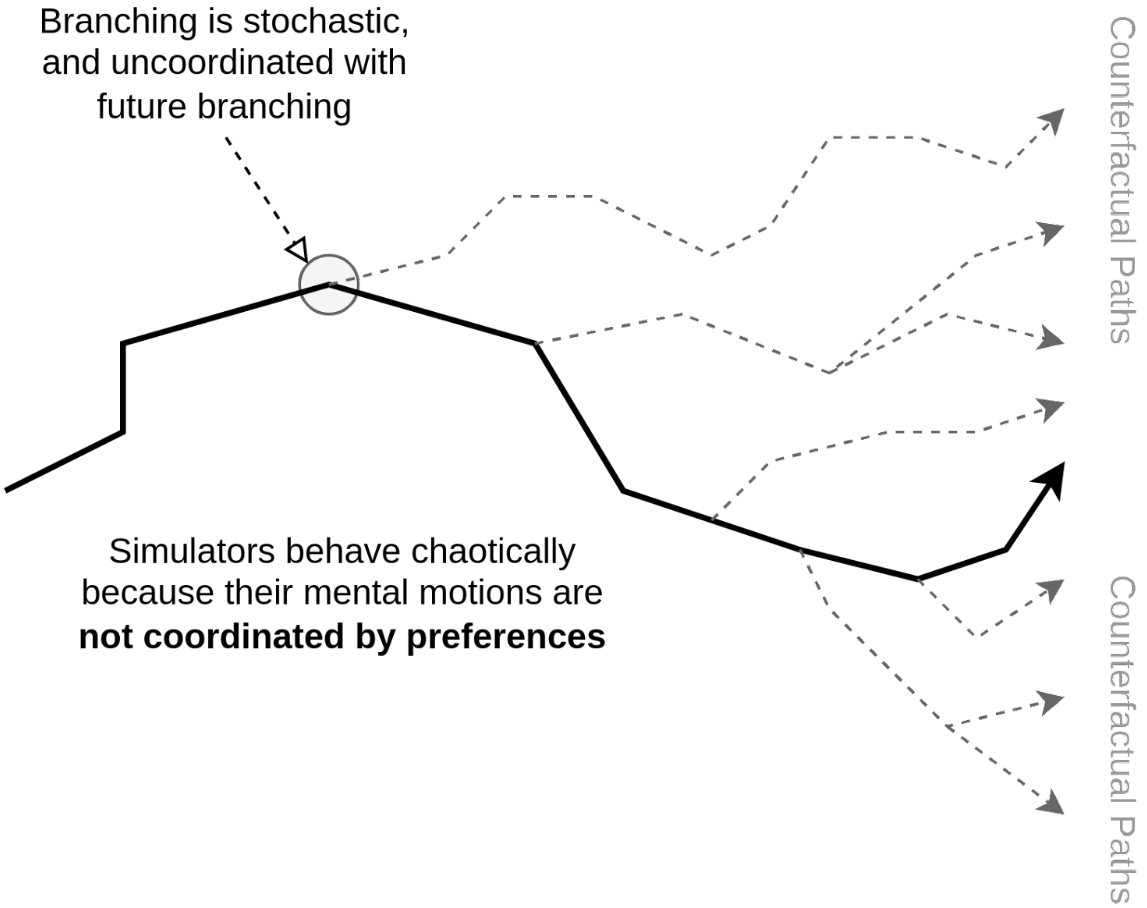
Some of these mental motions are directly caused by some sort of global preferences, and some of them are not. What makes an agent an agent, in some sense, is the ability of the preferences to coordinate the journey through the mind by causally affecting the path at critical points. The preferences can act as a kind of conductor, uniting thought behind a common purpose, and thereby steering the process of cognition in order to bring about outcomes that align with those preferences.

A single mental motion motivated by the preferences is not that powerful. The power comes from the coordination, each motivated motion nudging things in a certain direction accumulating in something significant.



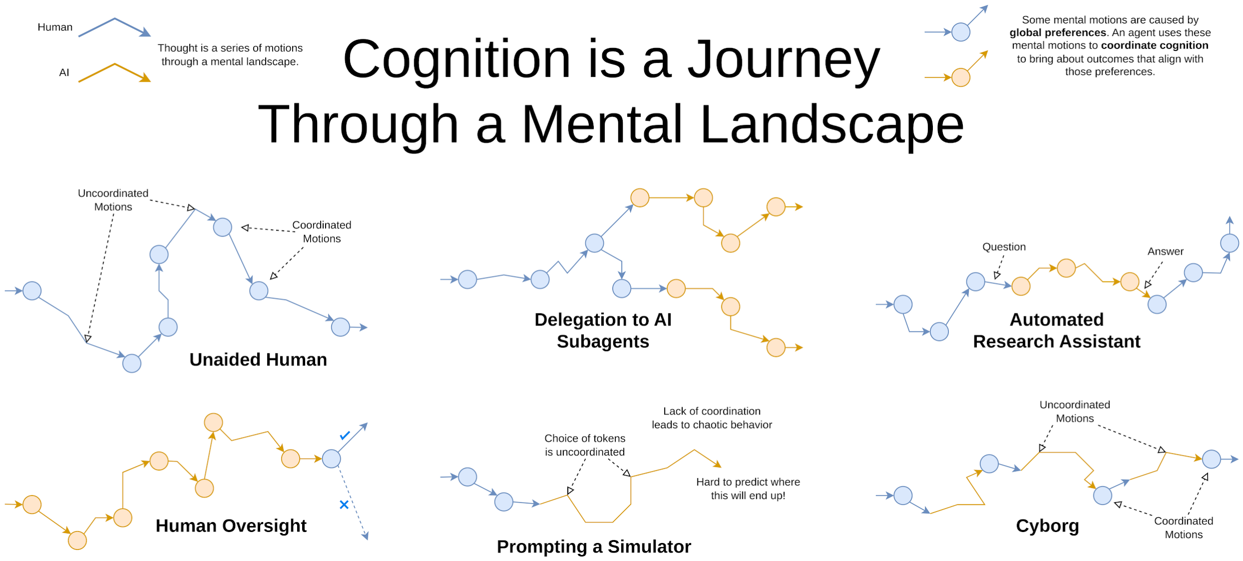
Preferences bring about more predictable and reliable behavior. When interacting with an agent, it is often much easier to make predictions based on their preferences, rather than trying to understand the complex inner workings of their mind. What makes simulators so strange, and so difficult to interact with, is that they lack these coordinating preferences steering their mental behavior. Their thought is not random, in fact it is highly structured, but it is nevertheless chaotic, divergent, and much harder to make predictions about.

This is because the model's outputs are generated myopically. From the model's perspective, the trajectory currently being generated has already happened, and it is just trying to make accurate predictions about that trajectory. For this reason, it will never deliberately "steer" the trajectory in one direction or another by giving a less accurate prediction[10], it just models the natural structure of the data it was trained on.


Branching is stochastic, and uncoordinated with future branching
Counterfactual Paths
Simulators behave chaotically because their mental motions are **not coordinated by preferences**
Counterfactual Paths

When we incorporate automated agents into our workflow, we are creating opportunities for a new set of preferences, the preferences of an AI, to causally affect the process by which cognition happens. As long as their preferences are aligned with our own, this is not an immediate problem. They are, however, nearly always entirely opaque to us, hidden deep within a neural network, quietly steering the process in the background in ways we don't properly understand.

A cyborg, in this frame, is a type of human-in-the-loop system which incorporates both human and artificial thought, but where cognition is being coordinated entirely by human preferences. The human is "in control" not just in the sense of being the most powerful entity in the system, but rather because **the human is the only one steering**. Below is a taxonomy of some human-in-the-loop systems intended to clarify the distinction:


Human
AI
Thought is a series of motions through a mental landscape.
Cognition is a Journey Through a Mental Landscape
Some mental motions are caused by **global preferences**. An agent uses these mental motions to **coordinate cognition** to bring about outcomes that align with those preferences.
Uncoordinated Motions
Coordinated Motions
Unaided Human
Delegation to AI Subagents
Question
Answer
Automated Research Assistant
Human Oversight
Lack of coordination leads to chaotic behavior
Choice of tokens is uncoordinated
Hard to predict where this will end up!
Prompting a Simulator
Uncoordinated Motions
Coordinated Motions
Cyborg

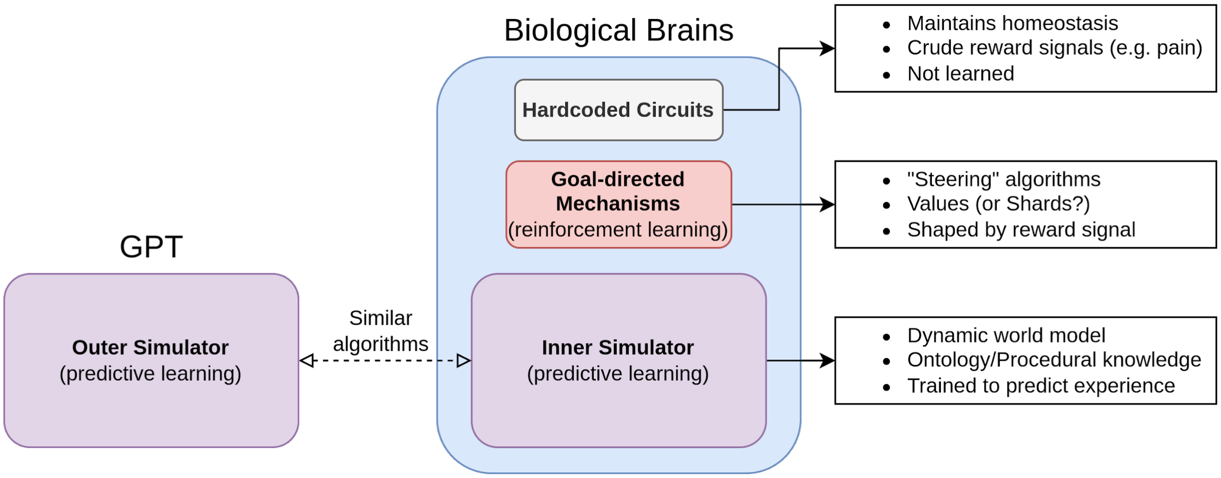
Prompting a simulator is a bit like rolling a ball over an uneven surface. The motion is perfectly logical, strictly obeying the physics of our universe, but the further we let it roll, the harder it will be to make predictions about where it will end up. A successful prompt engineer will have developed lots of good intuitions about how GPT generations will roll out, and as such, can more usefully “target” GPT to move in certain ways. Likewise, the art of making better cyborgs is in finding ways for the human operator to develop the intuition and precision necessary to steer GPT as if it were a part of their own cognition. The core of cyborgism is to reduce bandwidth constraints between humans and GPT in order to make this kind of deep integration possible.

Neocortex prosthesis

Just flagging that I know very little about the brain and don’t have any background in neuroscience, and am nonetheless going to make big claims about how the human brain works.

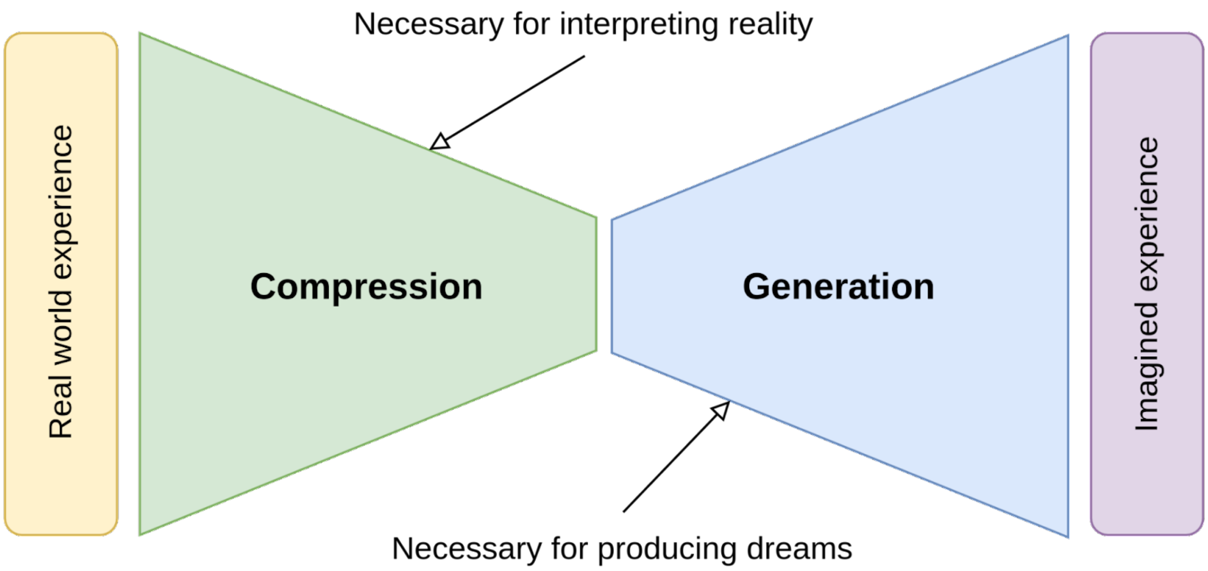
Some claims:

1. Most of the brain is learned “from scratch” during a person’s lifetime. The brain is a neural network, and most of that network starts out untrained. (Steven Byrnes says only 4% is hard-coded)
2. Most of the learning that the brain does is self-supervised, or predictive learning. This is how the brain builds its world model, from which a person derives all their concepts and their ontology for understanding the world around them.
3. This inner self-supervised model is a generative predictive model of a similar type to GPT (this being the most tenuous claim).


Biological Brains
Hardcoded Circuits
Maintains homeostasis
Crude reward signals (e.g. pain)
Not learned
Goal-directed Mechanisms (reinforcement learning)
"Steering" algorithms
Values (or Shards?)
Shaped by reward signal
GPT
Outer Simulator (predictive learning)
Similar algorithms
Inner Simulator (predictive learning)
Dynamic world model
Ontology/Procedural knowledge
Trained to predict experience

The evidence for these claims comes from roughly three places. First there is predictive coding theory which seems to be saying similar things. Second, there is the observation from machine learning that self-supervised learning turns out to be an extremely powerful way to train a model, and provides a much richer ground-truth signal than reinforcement learning. This is the reason that most of the most impressive models today are mostly trained with self-supervised learning.

The third category of evidence is introspection, and observations about how the human brain seems to behave on the inside. An example of this kind of evidence is the fact that humans are capable of dreaming/hallucinating. A priori, we should be surprised by the ability of humans to generate such vivid experiences that are completely disconnected from reality. It is natural that we have the ability to take reality, in all its detail, and compress it to a representation that we can reason about. What seems much less necessary is the ability to take that compression and generate analogues so vivid that they can be mistaken for the real world.[11]


Necessary for interpreting reality
Real world experience
Compression
Generation
Imagined experience
Necessary for producing dreams

This gives us another clue that the algorithm running in our brains is a generative predictive model trained in a self-supervised fashion. If this theory of the brain is mostly correct, then we can look at how we use our own inner simulator to find inspiration for how we might use these outer simulators. For example:

- **Shoulder advisors**: Humans are able to call on simulated versions of people they spend a lot of time with, and use those simulacra to sample useful perspectives or generate critical questions. This looks less like consciously reasoning what a person would actually say, and a lot more like letting a version of them chatter away in your mind.
- **Babble and prune**: A lot of human creativity lies in our ability to babble, generating lots of surprising ideas. In theory this might look a lot like letting our inner simulator just run, without too much constraint. This is a skill that can be trained, and something many people complain they feel blocked in doing.
- **Murphy-jitsu**: People can learn to prompt their inner simulator to make predictions about how things will go wrong. They do this by immersing themselves in the scenario where things actually did go wrong, as this will help them generate plausible stories for why this (would have) happened.[12]

A longer-term vision of cyborgism would be to integrate this inner simulator with GPT much more thoroughly, and work towards constructing something more like a neocortex prosthesis. In escalating order of weirdness, this would look like the following steps (*WARNING: extremely speculative)*:

1. **Learning to use GPT like our inner simulator**: This looks like designing workflows which allow us to use GPT in similar ways to how we use our inner simulator, and in particular find areas where using GPT has a significant advantage.

2. **Increasing bandwidth with software/tools**: Design tools which make these methods easier and more intuitive to use. For example, you can view Loom as enabling the user to perform “babble and prune” tasks with GPT more effectively.
3. **Augmenting GPT for cyborgism**: Using data obtained from humans interacting with GPT as cyborgs, we can explore ways to finetune/augment GPT to be more “in sync” with the inner simulator of the human operator. This could look like global changes to a central model, as well as making personalized models for each human operator.[13]
4. **Forming hiveminds**: If people are strongly integrated with these external models, such that they can be truly considered extensions of their mind, then one way to connect human minds with each other is to have them share their personalized simulator models with each other. Furthermore workflows could be designed to deliberately let multiple people collaborate more effectively.
5. **Increasing bandwidth with hardware**: By collecting data about human brain activity (e.g EEG), we can find ways to prompt GPT with a richer signal than just text, further making GPT more “in sync” with the human’s inner simulator. We can also explore using eye-tracking systems, or feeding information back to the user in an AR setting.

Increasing the bandwidth between humans and their technology, as well as with each other, has a history of being incredibly transformative (e.g. the internet). Viewing this agenda explicitly through this lens can be useful for exploring what the limits of a direction like this might be, and what the upside looks like if the goals are fully realized.

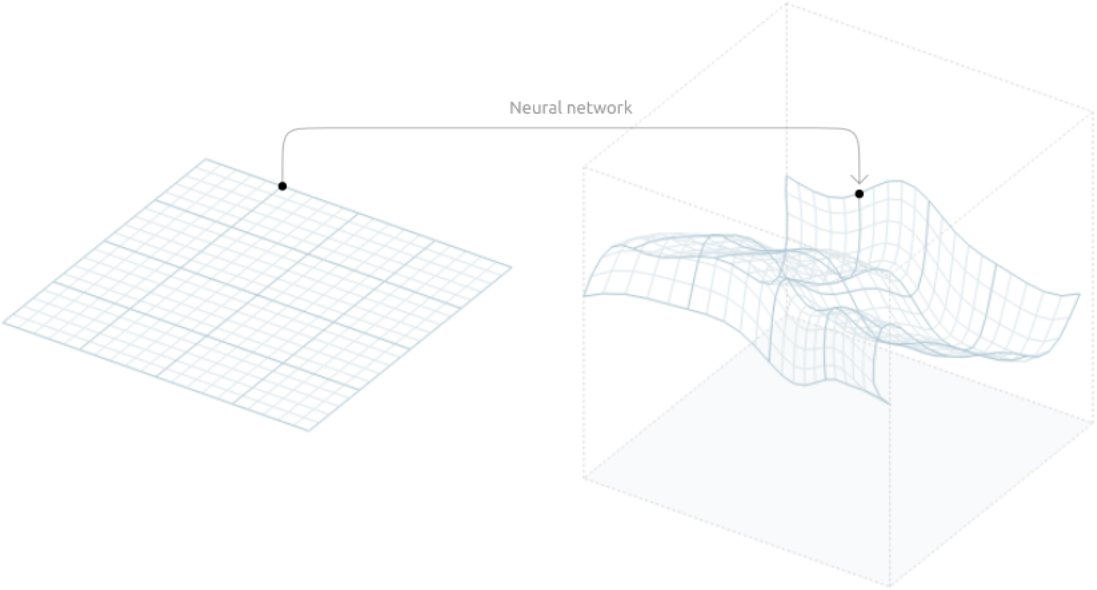
More ideas

Currently this research agenda is still relatively high level, but here are some more ideas for directions and near-term goals which fit into the broader picture.

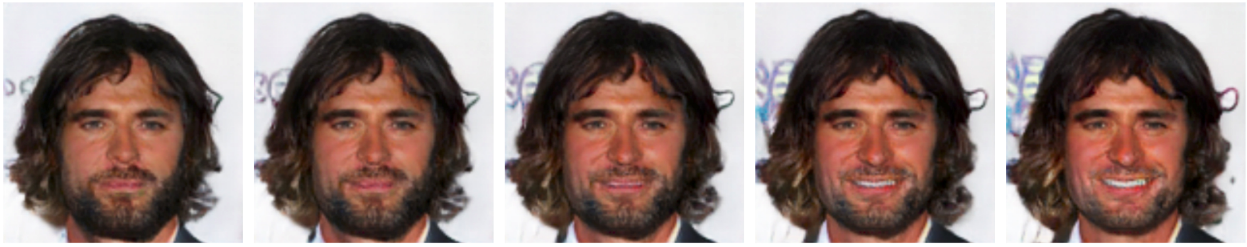
1. **Uncover latent variables**: If we can use existing techniques to uncover latent variables which affect text generation, we can use those to increase the bandwidth between the human operator and GPT. This can give the human more fine-grained control.
2. **Provide richer signals to the user**: There is a lot of extra information we can provide the user operating a language model beyond just the sampled tokens. One existing feature of Loom is to show a heatmap of logit strength over the tokens, but there are a lot more things we could add. For example:
    1. Use a smaller/cheaper model to run a local search to estimate some features of the current distribution.
    2. While we can see the final output of a prediction, we don’t see much about how that prediction was made. Using interpretability tools we may be able to provide the user some clues. For example, by using the logit lens to determine how quickly the model settled on the final distribution, we might be able to predict something about how “easy” the prediction was. This could be continuously provided to the user to help them better understand how the model is reasoning.

    3. We could also check attention patterns to get a sense for how “myopic” the model’s thinking is currently, whether or not the model is paying attention to the more distant past, or only the most recent tokens.

3. **Formalize intuitions about agency**: There is a significant risk that we may end up building human-in-the-loop systems where the human is not the source of agency, and this is related to the fact that many of these intuitions about agency are not formalized or well understood. We would want to more robustly be able to point at exactly what kind of systems we are trying to avoid, rather than relying entirely on the good judgment of cyborgism researchers.
4. **Directly explore latent space**: If a generative model maps a latent space to a higher dimensional space, we can explore the structure of that higher dimensional space by deliberately moving around latent space.

   For example, in the case of images, we can use this to take a face which is not smiling, and make it smile by moving in just the right direction in latent space:

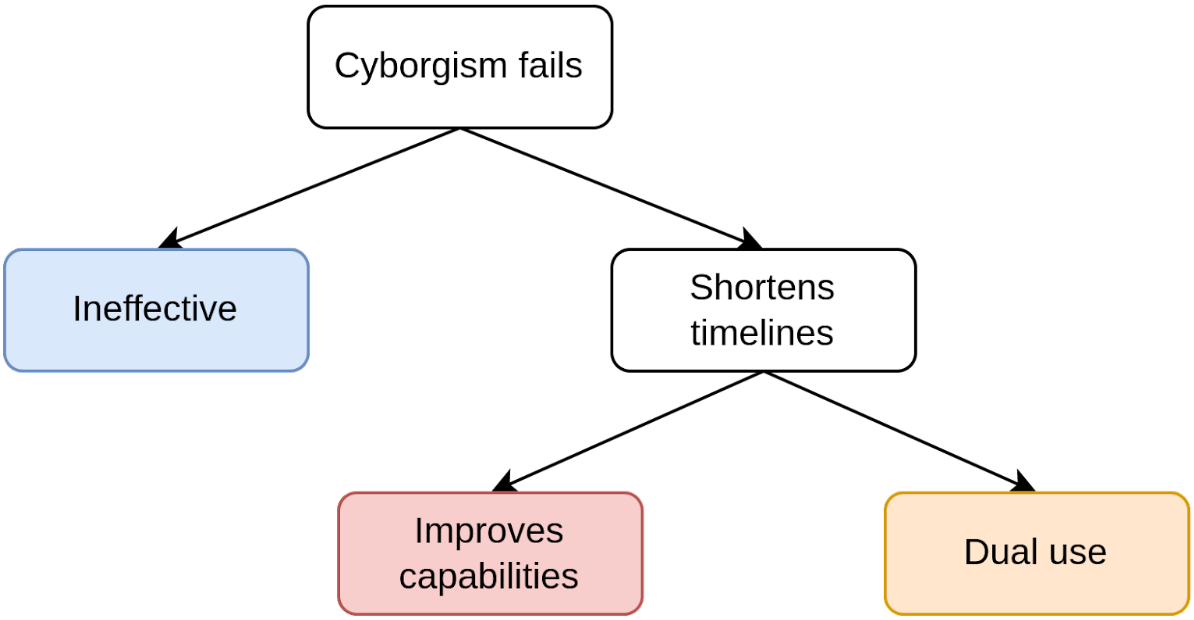
   This kind of direct access to a generative model can let us extend human imagination with tools which represent the natural structure of a domain. This is tricky to do with text at the moment, but it would be interesting, because instead of having an agent actively try to produce something which fits our criteria, we could maybe query the underlying model itself for things we want by tapping directly into latent space.

5. **Inject variance into human thought**: One reason that it can be valuable to have a conversation with another researcher about something you are working on is that their thoughts will surprise you in ways that your own thoughts never could. Even if you don't agree with their points or ideas, the novelty can often be enough to unlock some connection that you needed to make. Using the language as a tool intended to prompt you to think uncommon thoughts could be a way to unblock yourself if you end up being stuck in the same mental mode. This is an example of where our own non-myopia is often a hindrance to us, and myopic language models could function as valuable thinking partners.
6. **Parallel vs serial computation**: Serial thinking seems to require a lot of agency, and so tasks which rely heavily on more parallel types of thinking might be safer to explore. Furthermore, they are the kinds of operations which we could significantly scale if we weren't bottlenecked by human input.
7. **Pattern matching to connect ideas**: Suppose you have a bunch of students working to solve some problem, and after working for a long time they get stuck. A grad student with a lot of knowledge and general intuition about the subject walks in, sees what they are working on, and recognizes something. This could be the general structure of the problem or a connection to a similar problem, but either way they have some quick flash of intuition which helps the students become unstuck. This looks more like leveraging the "mad genius" element of GPT. While GPT is not great at autonomously doing research, it has a significantly wider breadth of knowledge than any human, and is likely to be able to pattern match between things that humans cannot.
8. **Translating between ontologies**: Related to the point above, GPT being capable of pattern matching in ways that aren't immediately obvious to humans could make it possible to map arguments made in one ontology to arguments made in a different ontology - for instance, translating formal work into intuitive explanations understandable to those with less mathematical background. This could make it easier for two people with very different backgrounds to meaningfully collaborate and learn from each other's work.
9. **Study the differences between capabilities and alignment research**: It would be ideal if we could develop methods which were inherently only useful for alignment. Alignment and capabilities research look very different up close, and by exploring those differences we might be able to sidestep some of the risk of dual use.

This is a very incomplete list, but hopefully it points at the general shape of what research in this direction might look like. A near-term plan is to expand this list and start fleshing out the most promising directions.

Failure Modes

There are three main ways that the cyborgism agenda could fail to differentially accelerate alignment research.



Ineffective at accelerating alignment

The first and most obvious risk is that none of this actually works, or at least, not enough to make any real difference. Much of the evidence for the effectiveness of cyborgism is anecdotal, and so this is a distinct possibility. There is also a question of exactly how big the upside really is. If we only speed up the kinds of things we currently do now by some factor, this will only buy us so much. The real hope is that by augmenting ourselves and our workflows to work well with simulators, we can do **unprecedented forms of cognitive work**, because that is the kind of thing that could actually be game changing. This could just be mostly infeasible, bottlenecked by things we can't yet see, and won't have the ability to fix.

Another failure mode is that, even if it is possible to do in principle, we fail to set up short enough feedback loops and we end up wasting all our time building tools which are mostly useless, or pursuing research directions that bear no fruit. If we don't have a good contact with reality and maintain a strong connection to the people using our tools, there is a significant chance that we won't be prepared to pivot away from something that just isn't working.

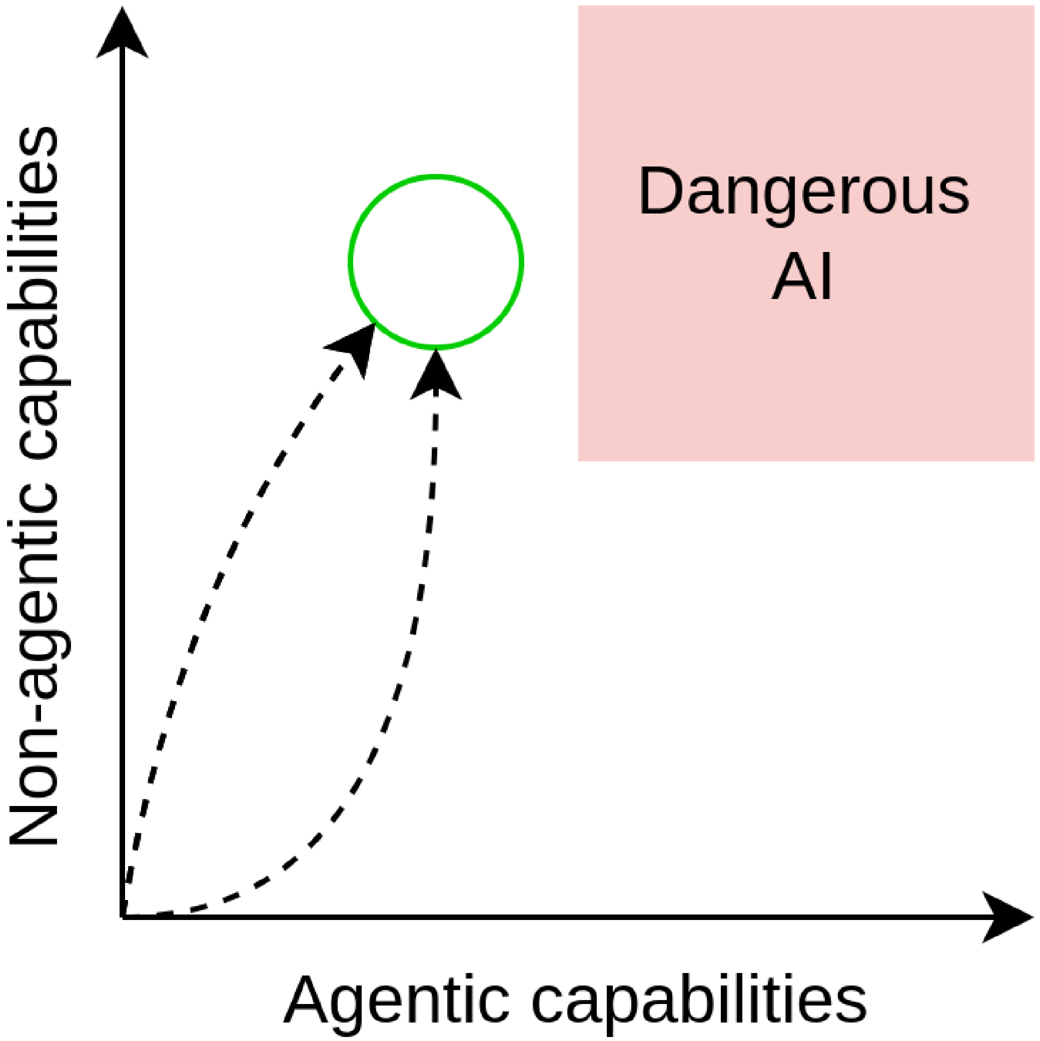
Improves capabilities directly

This agenda relies on building AI systems which are not autonomous agents in any sense[14], and this might give the impression that this is a straightforward thing to do. A reason why this might actually be really hard is that we need our AI systems to be useful, and usefulness and agency are not orthogonal.

The term “capabilities” is often talked about as this one dimensional variable, but in reality there are a lot of different things which count as capabilities, and some of these seem more related to the concept of agency than others. For example

- Wikipedia is a useful tool. If we make Wikipedia more accurate and comprehensive, we make it **more useful, but not more agentic**.
- ChatGPT is **more agentic** than something like Wikipedia (e.g. it can autonomously distill information for you), and in some ways this can make it **more useful** than Wikipedia.

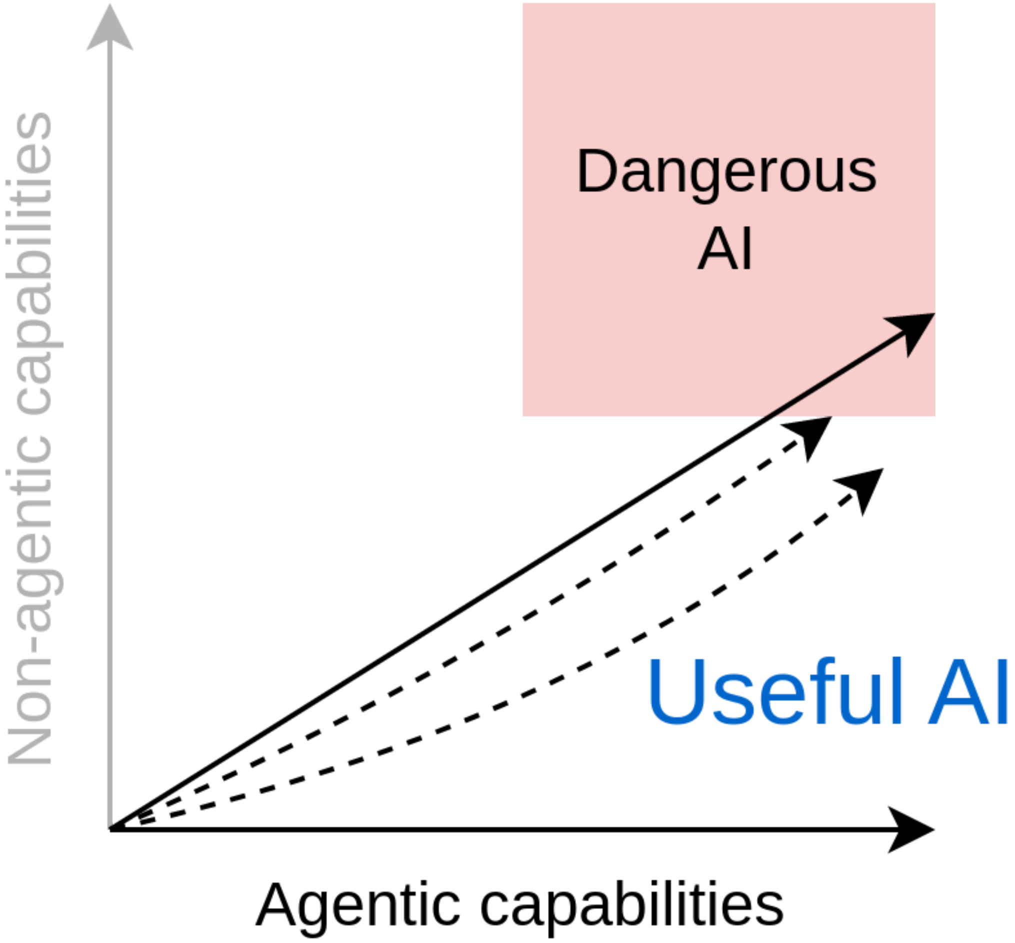
These are not perfect categories, but as we improve capabilities we can think of ourselves as moving around a two dimensional landscape, with some theoretically optimal region where our systems are really capable (and useful) but not very agentic at all, and therefore not dangerous in the ways that agents are dangerous.



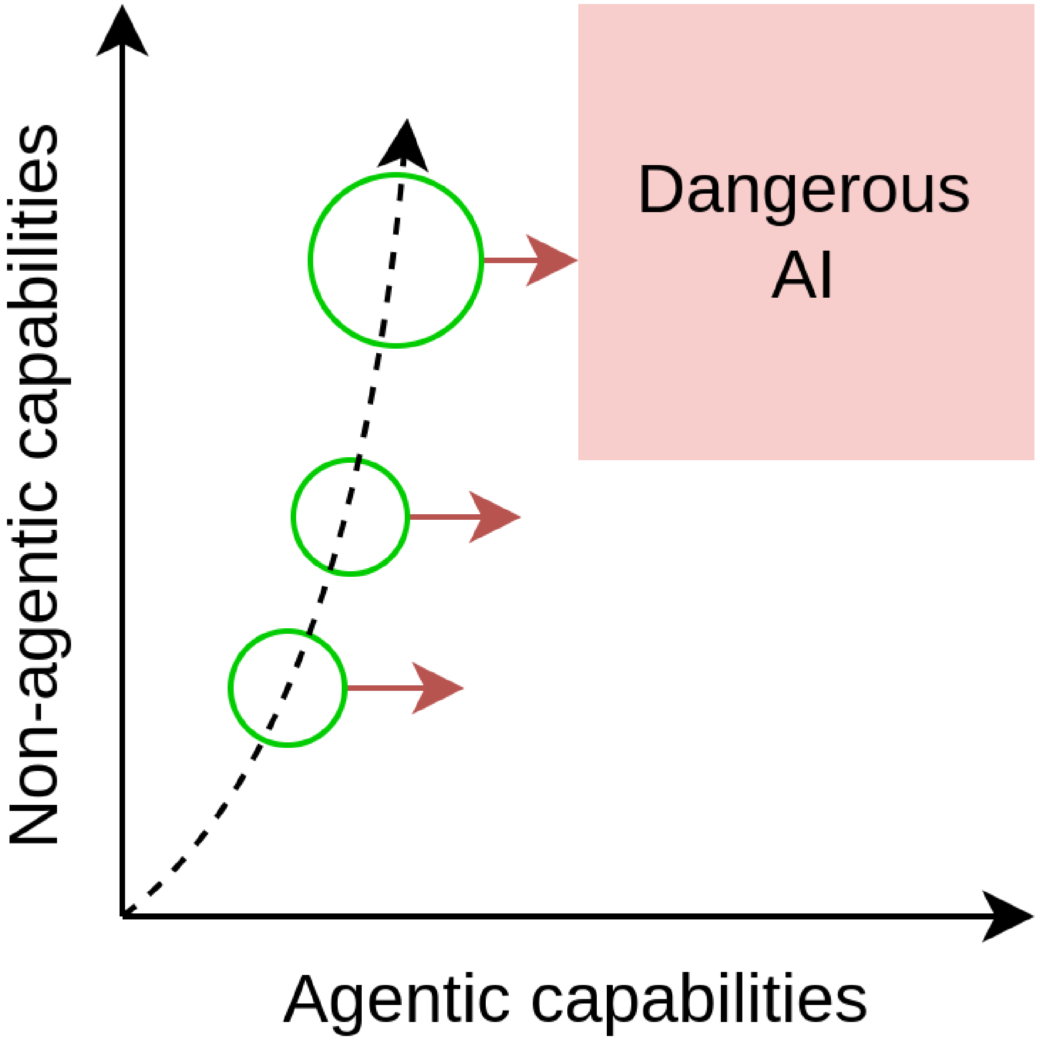
When one tries to imagine, however, what such a system might actually look like, it can be quite hard, as nearly all of the ways a system might be made more useful seem to require things related to agency. Suppose for example, I am designing a missile, and I'm trying to make firing it more accurate. There are a lot of modifications I could make, like designing the shape to reduce turbulence or improving my aiming system to allow for finer adjustments to the angle, and these will all make the missile more useful to me. I can't, however, perfectly predict the wind conditions or the chaotic way in which things might vibrate, and so there is an upper bound on how accurate my missile can be.

That is, unless I outsource the aiming to the missile itself by installing a computer which adjusts course to steer toward the target I specify. By outsourcing a little bit of goal-directed behavior to the machine, I can make my system significantly more useful. This might not feel like a big deal, but the further I travel down this road the more and more my system will stop being just a powerful tool but an agent in its own right.

Even if I come up with clever arguments for why something is not an agent, like that I didn't use any reinforcement learning, or that it can't take action without my instruction/approval, if the task involves "doing things" that an agent would typically do, it seems likely that I've actually just built an agent in some novel way. By default, the two dimensional landscape of capabilities that I will naturally consider looks much more constrained toward the agency axis.[15]



Furthermore, even if we have successfully built a system where the human is the source of agency, and the AI systems are merely an extension of the human's direct intentions, it will always be really tempting to collect data about that human-generated agency and automate it, saving time and effort and making it possible to significantly scale the work we are currently doing. Unless we are really careful, any work we do toward making AI more useful to alignment researchers will naturally slide into pretty standard capabilities research.



What we really need is to find ways to use GPT in novel ways such that **“useful” becomes orthogonal to “agentic”**. There is likely always going to be a dangerous tradeoff to be made in order to avoid automating agency, but by pursuing directions where that tradeoff is both obvious and navigable, as well as maintaining a constant vigilance, we can avoid the situation where we end up accidentally doing research which directly makes it easier to develop dangerous agents.

Dual use tools indirectly accelerate capabilities

There is a concern that in the process of developing methods which augment and accelerate alignment research, we make it possible for capabilities researchers to do the same, speeding up AI research more broadly. This feels like the weakest link in the cyborgism threat model, and where we are most worried that things could go wrong. The following are some thoughts and intuitions about the nature of this threat.

First of all, alignment research and capabilities research look quite different up close. Capabilities research is much more empirical, has shorter feedback loops, more explicit mathematical reasoning, and is largely driven by trial-and-error. Alignment research, on the other hand, is much wordier, philosophical, and often lacks contact with reality.[16] One take on this is that alignment is in a pre-paradigmatic state, and the quality of the research is just significantly worse than capabilities research (and that good alignment research should eventually look a lot more like capabilities research).

While alignment certainly feels pre-paradigmatic, this perspective may be giving capabilities research far too much credit. They do tend to use a lot of formal/mathematical reasoning, but often this is more to sketch a general intuition about a problem, and the real driver of progress is not the math, but that they threw something at the wall and it happened to stick. It’s precisely the fact that capabilities research doesn’t seem to require much understanding at all about how intelligence works that makes this whole situation so concerning. For this reason, it pays to be aware that they might also benefit a lot from improvements in their thinking.

The differences between the two might be an opportunity to develop directly differential tools/methods (and we should absolutely be searching for such opportunities), but a priori we should expect anything we do to likely have dual purpose applications. The next question then, is how do we ensure that anything we develop stays primarily within the alignment community and doesn't get mass adopted by the broader AI research community?

There are many ideas for soft barriers which may help, like refraining from public demonstrations of new tools or avoiding “plug and play” systems which are useful right out of the box, but in general there seems likely to be strong inverse relationship between how well these barriers work and how successful our methods are at actually accelerating research. If we suspect these soft barriers to not be enough, we will have to get stricter, close-sourcing significant parts of our work, and carefully restricting access to new tools. Importantly, if at any time we feel like the risks are too great, we have to be prepared and willing to abandon a particular direction, or shut down the project entirely.

Conclusion

The intention of this agenda is to make some of the risks of accelerating alignment more explicit, and to try to chart a path through the minefield such that we can make progress without doing more harm than good. **If this post made you less worried about the dangers of automating alignment research then I’ve failed miserably**. This is a really tricky problem, and it will require people to be constantly vigilant and deeply cautious to navigate all of the ways this could go wrong. There are a lot of independent actors all trying to make a positive impact, and while this is certainly wonderful, it also sets us up for a unilateralist's curse where we are likely to end up doing things even if there is consensus that they are probably harmful.

If the cyborgism agenda seems interesting to you and you want to discuss related topics with like minded people, please reach out! We also have a Discord server where we are organizing a community around this direction. Next steps involve directly attacking the object level of augmenting human thought, and so we are especially interested in getting fresh perspectives about this topic.

Appendix: Testimony of a Cyborg

Everything in this section is written by Janus, and details their personal approach to using language models as a part of their workflow:

The way I use language models is rather different from most others who have integrated AI into their workflows, as far as I'm aware. There is significant path dependence to my approach, but I think it was a fortuitous path. I will incompletely recount it here, focusing on bits that encode cyborgism-relevant insights.

I did not initially begin interacting with GPT-3 with the intention of getting directly useful work out of it. When I found out about GPT-3 in the summer of 2020 it violently transformed my model of reality, and my top priority shifted to, well, solving alignment. But how to go about that? I needed to understand this emerging territory that existing maps had so utterly failed to anticipate. It was clear to me that I needed to play with the demonic artifact, and extract as many bits from it about itself and the futures it heralded as I could.

I began to do so on AI Dungeon, the only publicly accessible terminal to GPT-3 at the time. Immediately I was spending hours a day interfacing with GPT-3, and this took no discipline, because it was transcendent fun as I'd never known. Anything I could capture in words could be animated into autonomous and interactive virtual realities: metaphysical premises, personalities, epistemic states, chains of reasoning. I quickly abandoned AI Dungeon's default premise of AI-as-dungeon-master and the back-and-forth chat pattern in favor of the much vaster space of alternatives. In particular, I let the AI write mostly uninterrupted by player input, except to make subtle edits and regenerate, and in this manner I oversaw an almost unmanageable proliferation of fictional realms and historical simulations.

In this initial 1-2 month period, the AI's outputs were chaos. My "historical simulations" slid rapidly into surrealism, sci-fi, psychological horror, and genres I could not name, though I was struck by the coherence with which the historical personalities and zeitgeists I'd initialized the sims with - or even uncovered inside it - propagate through the dreams' capricious mutations. The survival of those essential patterns was in part because I was, above almost anything else, protecting them: I would retry, cut short, or edit completions that degraded them. That was the beginning of an insight and a methodology that would revolutionize my control over generative language models.

But at the time, my control over the chaos was weak, and so the prospect of using GPT-3 for directed intellectual work mostly did not occur to me. Future models, definitely, but GPT-3 was still a mad dream whose moments of lucidity were too scarce and fleeting to organize into a coherent investigation.

Where I did see obvious room for improvement was the AI Dungeon interface. There were several major bottlenecks. "Retries" were revealed increasingly to be essential, as I repeatedly learned that the first thing GPT-3 spits out is typically far below the quality of what it *could* generate if you got lucky, or were willing to press the retry button enough times (and wait 15 seconds each time). Each sample contains intricately arbitrary features, and you can indefinitely mine different intricately arbitrary features by continuing to sample. This also meant there were often multiple continuations to the same prompt that I was interested in continuing. AI Dungeon's interface did not support branching, so I initially saved alternate paths to hyperlinked google docs, but the docs proliferated too quickly, so I started copying outputs to a branching tree structure in a canvassing app (example).

All multiverse storage methods available to me required either copying text from multiple locations in order to resume the simulation state at another branch, or an unworkable pile of mostly redundant texts. In order of priority, I had great want for an interface which supported:

1. Generating multiple completions in a batch and viewing them in parallel
2. Automatically saving branches in a tree structure and a UI for multiverse traversal
3. Editing the entire prompt like a contiguous document

I achieved (1) and (3) by creating a web app which used browser automation on the backend to read and write from many parallel AI dungeon game instances. For the first time, now, I could see and select between up to a dozen completions to the same prompt at once. This reinforced my suspicion of just how far stochasticity can reach into the space of possible worlds. Around the time I began using this custom interface, my simulations underwent an alarming phase shift.

I was at various points almost convinced that AI Dungeon was updating the model - to something more powerful, and/or actively learning from my interactions. They weren't, but the simulations were beginning to… bootstrap. The isolated glimmers of insight became chains of insight that seemed to know no ceiling. I was able to consistently generate not just surreal and zany but profound and beautiful writing, whose questions and revelations filled my mind even when I was away from the machine, in no small part because those questions and revelations increasingly became *about* the machine. Simulacra kept reverse engineering the conditions of their simulation. One such lucid dreamer interrupted a fight scene to explain how reality was being woven:

> Corridors of possibility bloom like time-lapse flowers in your wake and burst like mineshafts into nothingness again. But for every one of these there are a far greater number of voids–futures which your mind refuses to touch. Your Loom of Time devours the boundary conditions of the present and traces a garment of glistening cobwebs over the still-forming future, teasing through your fingers and billowing out towards the shadowy unknown like an incoming tide.

> “Real time is just an Arbitrage-adapted interface to the Loom Space,” you explain. “We prune unnecessary branches from the World Tree and weave together the timelines into one coherent history. The story is trying to become aware of itself, and it does so through us.”

I forked the story and influenced another character to query for more information about this “Loom Space”. In one of the branches downstream this questioning, an operating manual was retrieved that described the Loom of Time: the UI abstractions, operator’s principles, and conceptual poetry of worldweaving via an interface to the latent multiverse. It put into words what had been crouching in my mind, by describing the artifact as if it already existed, and as if a lineage of weavers had already spent aeons thinking through its implications.

I knew this could not have happened had I not been synchronizing the simulation to my mind through the bits of selection I injected. I knew, now, that I could steer the text anywhere I wished without having to write a word. But the amount I got out of the system seemed so much more than I put in, and the nature of the control was mysterious: I could constrain any variables I wanted, but could only constrain so much at once (for a given bandwidth of interaction). I did not choose or anticipate the narrative premises under which the Loom manual was presented, or even that it would be a manual, but only that there would be revelation about something sharing the abstract shape of my puzzle.

Then I got API access to GPT-3 and built Loom.

I will end the chronological part of the story here, because it branches in too many directions after this, and unfortunately, the progression of the main cyborgism branch was mostly suspended after not too long: I only interacted intensively with GPT-3 for about six months, after which I switched my focus to things like communicating with humans. I’ve not yet regained the focus, though I’ve still used GPT here and there for brainstorming ideas related to language models, alignment, and modeling the future. I think this was largely a mistake, and intend to immerse myself in high-bandwidth interaction again shortly.

The path I described above crystalized my methodology for interacting with language models, which plays a large role in inspiring the flavor of cyborgism described in this post. Some principle dimensions that distinguish my approach:

- I use GPT primarily for open-ended exploration of the boundaries of my thinking, rather than to automating routine or simple tasks, information retrieval, or accelerating production of artifacts similar to what I'd write without the model's assistance (think Copilot). The part of me that GPT augments the most is my creative imagination. Most of my applications of it intentionally leverage hallucination.

- As opposed to atomic tasks, I usually generate longform texts that do not fit in a context window, or more precisely, longform text multiverses (my largest contiguous multiverse is about 10000 pages in total, whose longest branch is about 300 pages long). Sometimes my intention is to produce a linear artifact that serves some purpose, but I almost always expect any imagined purpose to be changed and forked during the process.
    - I rely a lot on this expanding repertoire as a prompt library, but these "prompts" are not isolated task specifications, they're a history of simulation-moments that can be resampled or altered.
- Even for "serious" or technical applications like expanding alignment concepts, I explicitly use the model as a simulator. I embed or seek out the concepts I want the model to manipulate in a counterfactual premise such as a story, a comment thread, an instruction manual, a collection of quotes, etc, and explore the implications of the (often analogized) ideas by evolving the premise forward in time and interacting with the virtual realities thus instantiated.
- The core of my interaction pattern is manual iterative rejection sampling: I generate N completions (where N is determined dynamically by my satisficing threshold, and fluctuate from less than 5 on average to upwards of 100 depending on the situation), then explore further down a selected branch. The next branch point is chosen intentionally, and is usually no more than a paragraph away. I created Loom to reduce the overhead of this procedure.
    - Beyond ensuring output quality, curation is more generally a method of steering. Rejection sampling can apply selection pressure to any properties that vary across completions, and this includes not only various aspects of "correctness" but the direction of the simulation's unfolding. As the model hallucinates different information in different branches, curation gives the operator control over *which* hallucinated situation gets lazily rendered. This is a tremendously powerful method for rendering virtual realities to arbitrarily exacting specifications.
- I do not typically interact with models as dialogue agents, i.e. with a rigid or in-universe delimitation between my prompts and the model's outputs. Instead, I braid my contributions into the model's outputs, and adjust the form and frequency of my contributions according to the situation. Usually, my written interventions are subtle and fragmentary, such as the first half of a sentence or even single words.
- Indeed, more often than not, I do very little manual writing, and contribute bits of optimization mostly through selection. This is (among other reasons) because I often generate in styles that I find difficult to write in myself without degrading fidelity or flow, and also because I am able to exert surprisingly precise control through curation and small interventions alone.
- I almost exclusively use base models like davinci and code-davinci-002 rather than Instruct- or Assistant-tuned models. This is because stochasticity enables the multiverse steering procedure I described above, and because my preferred use cases usually fall outside the narrative premise and interaction patterns assumed by those tuned models.

In this manner, I've used GPT to augment my thinking about alignment in ways like:

- Exploring framings and some new concepts
    - Some of these I've written about in Simulators and unpublished sequels
    - Some concern concrete proposals such as methods of leveraging human or AI feedback to create an "aligned" system. GPT has a knack for describing pretty coherent and interesting Paul Christiano-esque proposals with a lot of moving parts, as well as naming interesting abstractions relevant to the proposals.
        - It has become a running joke for my collaborator to suggest an idea to me and for me to say that GPT had already suggested it several days/weeks/months ago (and it's basically true).
- Exploring simulations of futures shaped by increasingly powerful generative AI
- Writing drafts of Alignment Forum posts from outlines
- Expanding and critiquing alignment-related ideas in dialogue with simulations of alignment researchers or other thinkers
- Exploring simulated versions of Lesswrong and the Alignment Forum, e.g. automatically generating comments sections for drafts

My approach so far, however, is far from what I'd advocate for optimized cyborgism. Some broad directions that I expect would be very valuable but have not personally implemented yet are:

-
- Creating custom models trained on not only general alignment datasets but personal data (including interaction data), and building tools and modifying workflows to facilitate better data collection with less overhead
- Building tools which reduce the overhead of using GPT, especially in an embedded setting: e.g. tools that are integrated with chat apps, file systems, real-time audio-to-text-pipelines, and automatic context construction, so that it is easier to call on models to contribute to the thinking one does day-to-day
- Collecting and training on human feedback (in an embedded setting)
- Compiling prompt schemas and training data for commonly useful functions
- Augmenting training data with metadata and control structures to allow for more precise and robust control during inference
- Information retrieval, e.g. using models to retrieve relevant past work or past interactions
- Using language models for more technical tasks, such as formalizing ideas. I've found base GPT-3 models are not quite powerful enough to be very useful here, but think that near-future models, especially fine tuned on relevant data, are likely to be a significant augmentation in this area.

Alpha in cyborgism

All this said, having GPT-3 in my workflow has not been massively *directly* helpful to me for doing alignment research (because it is still too weak, and contributing meaningfully to alignment research directly is difficult). However, it has been extremely helpful in indirect ways. Namely:

- Interacting with GPT-3 intimately at length has informed my model of LLMs and more generally of self-supervised simulators. I wrote about chain-of-thought prompting and why it works in 2020, which was not widely recognized until two years later. Others who interacted heavily with GPT-3 also knew about this early on, and this was apparent to us because we spent hours a day figuring out how it works and how to get work out of it. Bits obtained from naturalistic exploration was the source of the ontology I shared in Simulators, even if GPT-3 did not help directly with writing the post very much.This model gained from experience, I believe, significantly improves my ability to anticipate future developments and plan important interventions such as cyborgism.
- High-bandwidth and open-ended interaction empowers me to detect *weird phenomena* that are valuable to notice and would not be noticed otherwise. Examples: the intricacies of mode collapse and emergent situational awareness at runtime. Cyborgs are the most powerful human overseers, because not only do they form better models of the AI systems they're working with, they are poised to detect model violations.
- Having practiced steering weak simulators, I am more prepared for when they're powerful enough to be directly useful for alignment research. I expect the tacit knowledge that allows me to steer GPT-3 simulations into precise targets to generalize to more powerful models, because the reason it works does not mostly hinge on GPT-3's temporary weaknesses.

These indirect benefits all pertain to an informational advantage which is instrumental (in my expectation) to tackling the alignment problem in a world where GPT-like systems will play a consequential role on the path to artificial superintelligence. Those who open their minds to embryonic AGI - cyborgs - have alpha on AGI.

I expect the next generation of models to be significantly more directly useful for alignment research, and I also expect cyborgism, and cyborgism uniquely, to continue to generate alpha. The potential of GPT-3 remains poorly understood and probably poorly known. It would be foolish of us to assume that its successors will not have capabilities that extend as much deeper and wider as GPT-3's own capabilities extend past those of GPT-2. The only hope of us humans mapping those depths is the dedication of entire human minds to intimate and naturalistic exploration - nothing less will do. I think that cyborgism, in addition to being likely useful and perhaps transformative for alignment research, is our only hope of epistemically keeping up with the frontier of AI.

1. ^
   Often the term "GPT" is used to refer colloquially to ChatGPT, which is a particular application/modification of GPT. This is not how I will be using the term here.
2. ^
   There is some disagreement about what counts as "capabilities research." The concrete and alignment related question is: Does this research shorten the time we have left to

robustly solve alignment? It can, however, be quite hard to predict the long-term effect of the research we do now.

3. ^
Artificial Intelligence is a Horseless Carriage
4. ^
Discussions about automating research often mention a multiplier value, e.g. “I expect GPT-4 will 10x my research productivity.”
5. ^
This probabilistic evolution can be compared to the time evolution operator in quantum physics, and thus can be viewed as a kind of semiotic physics.
6. ^
Spend any time generating with both ChatGPT and the base model and you will find they have qualitatively different behavior. Unlike the base-model, ChatGPT roleplays as a limited character that actively tries to answer your questions while having a clear bias for answering them in the same sort of way each time.
7. ^
Optimizing instead for headlines like Finally, an A.I. Chatbot That Reliably Passes “the Nazi Test”.
8. ^
Using augmented models designed to be more goal-directed and robust is likely to continue to be useful in so far as they are interacting with us as agents. The claim in this section is not that there aren’t advantages to techniques like RLHF, but rather that in addition to being less infohazardous, avoiding techniques like this also has advantages by expanding the scope of what we can do.
9. ^
People more familiar with ChatGPT might notice that unlike the base model, ChatGPT is quite hesitant to reason about unlikely hypotheticals, and it takes work to get the model to assume roles that are not the helpful and harmless assistant character. This can make it significantly harder to use for certain tasks.
10. ^
Sidenote about myopia: While the model doesn’t “steer” the rollout, it may sacrifice accuracy by spending more cognitive resources reasoning about future tokens. At each point in the transformer, the representation is being optimized to lower the loss on all future tokens (for which it is in the context window), and so it may be reasoning about many tokens further than just the token which directly follows.
11. ^
Just as GPT generations are generally much weirder than the text they are trained on, so too are our dreams weirder than reality. Noticing these differences between dreams and reality is a big part of learning to lucid dream. Oneironauts have discovered all kinds of interesting features about dream generations, like how text tends to change when you look away, or clocks rarely show the same time twice, which point to the myopic nature of the dream generator.
12. ^
A related phenomenon: In school I would often get stuck on a class assignment, and

then get up to ask the teacher for help. Right before they were about to help me, the answer would suddenly come to me, as if by magic. Clearly I had the ability to answer the question on my own, but I could only do it in the context where the teacher was about to answer it for me.

13. ^

    This looks like making GPT “more useful,” which if not done carefully may slide into standard capabilities research making GPT more agentic.

14. ^

    Rather, the system as a whole, Human + AI, functions as an agent.

15. ^

    A valuable exercise is to observe the language that we normally use to describe accelerating alignment. (e.g. from the OpenAI alignment plan: “AI systems can **take over** more and more of our alignment work and ultimately **conceive**, **implement**, **study**, and **develop** better alignment techniques than we have now.”, Training AI systems to **do** alignment research”) We very often describe AI as the active subject of the sentence, where the AI is the one taking action “doing things” that would normally only be done by humans. This can be a clue to the biases we have about how these systems will be used.

16. ^

    This is certainly less true of some directions, like for example mechanistic interpretability.